Lecture 3: Multiple Regression

Prof. Sharyn O'Halloran

Sustainable Development U9611

Econometrics II

Outline

- Basics of Multiple Regression
  - Dummy Variables
  - Interactive terms
  - Curvilinear models
- Review Strategies for Data Analysis
  - Demonstrate the importance of inspecting, checking and verifying your data before accepting the results of your analysis.
  - Suggest that regression analysis can be misleading without probing data, which could reveal relationships that a casual analysis could overlook.
- Examples of Data Exploration

Multiple Regression

Data:

Y	X_1	X_2	X_3
34	15	-37	3.331
24	18	59	1.111
…	…	…	…

Linear regression models (Sect. 9.2.1)

1. Model with 2 X's: $\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
2. Ex: Y: 1st year GPA, X_1: Math SAT, X_1:Verbal SAT
3. Ex: Y= log(tree volume), X_1:log(height), X_2: log(diameter)

Important notes about interpretation of β's

- Geometrically, $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ describes a plane:
  - For a fixed value of X_1 the mean of Y changes by β_2 for each one-unit increase in X_2
  - If Y is expressed in logs, then Y changes β_2% for each one-unit increase in X_2, etc.
- The meaning of a coefficient depends on which explanatory variables are included!
  - β_1 in $\mu(Y|X_1) = \beta_0 + \beta_1 X_1$ is not the same as
  - β_1 in $\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Specially constructed explanatory variables

- ***Polynomial terms***, e.g. X^2, for curvature (see Display 9.6)
- ***Indicator variables*** to model effects of categorical variables
  - One indicator variable (X=0,1) to distinguish 2 groups;
    - Ex: X=1 for females, 0 for males
  - (K-1) indicator variables to distinguish K groups;
    - Example:
      - X_2 = 1 if fertilizer B was used, 0 if A or C was used
      - X_3 = 1 if fertilizer C was used, 0 if A or B was used
- ***Product terms*** for interaction

$$\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$$

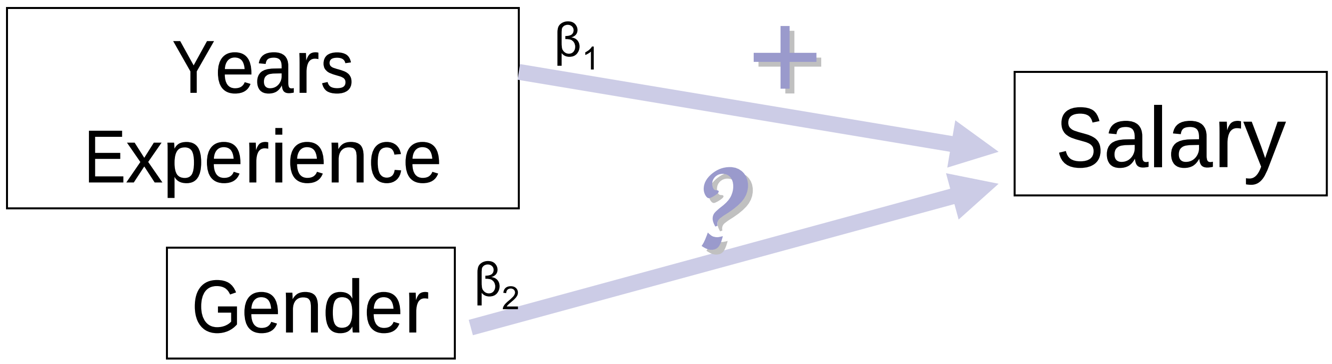
$\rightarrow$ $\mu(Y|X_1,X_2=7) = (\beta_0 + 7\beta_2) + (\beta_1 + 7\beta_3)\, X_1$

$\mu(Y|X_1,X_2=-9) = (\beta_0 - 9\beta_2) + (\beta_1 - 9\beta_3)\, X_1$

"The effect of X_1 on Y depends on the level of X_2"

Sex discrimination?

- Observation:
  - Disparity in salaries between males and females.
- Theory:
  - Salary is related to years of experience
- Hypothesis
  - If no discrimination, gender should not matter
  - Null Hypothesis $H_0 : \beta_2 = 0$



Hypothetical sex discrimination example

Data:

Y_i = salary for teacher i,

X_{1i} = their years of experience,

X_{2i} = 1 for male teachers, 0 if they were a female

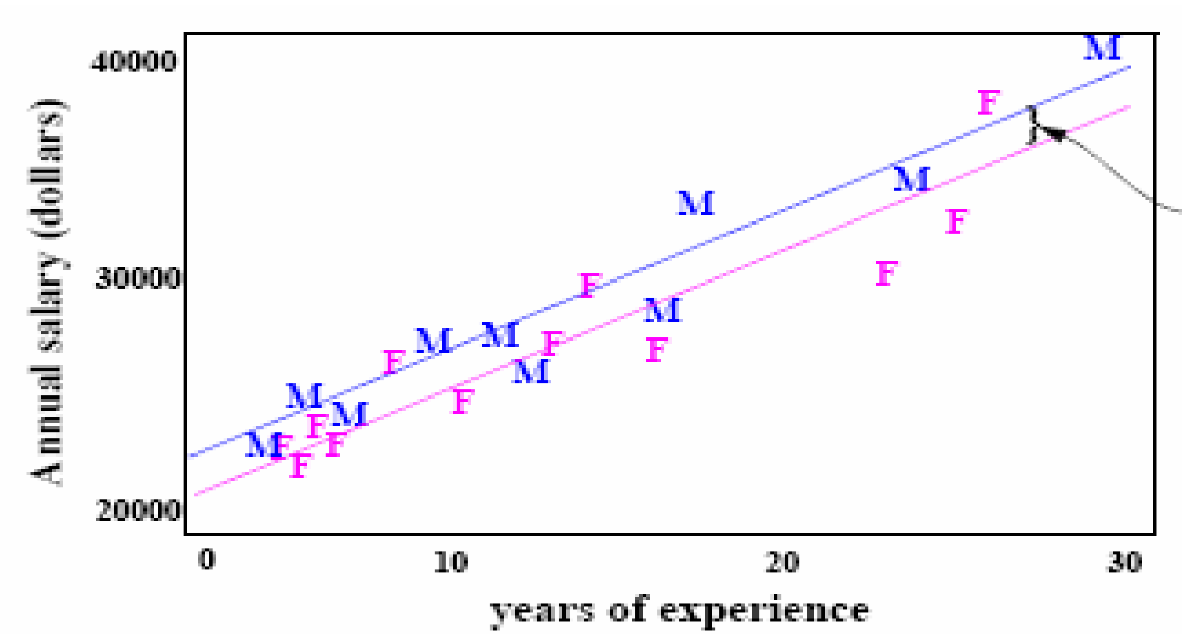
i	Y	X_1	Gender	X_2
1	23000	4	male	1
2	39000	30	female	0
3	29000	17	female	0
4	25000	7	male	1

"Gender": Categorical factor

X_2 Indicator variable

Model with Categorical Variables

- Parallel lines model: $\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - for all females: $\mu(Y|X_1,X_2=0) = \beta_0 + \beta_1 X_1$
  - for all males: $\mu(Y|X_1,X_2=1) = \beta_0 + \beta_1 X_1 + \beta_2$



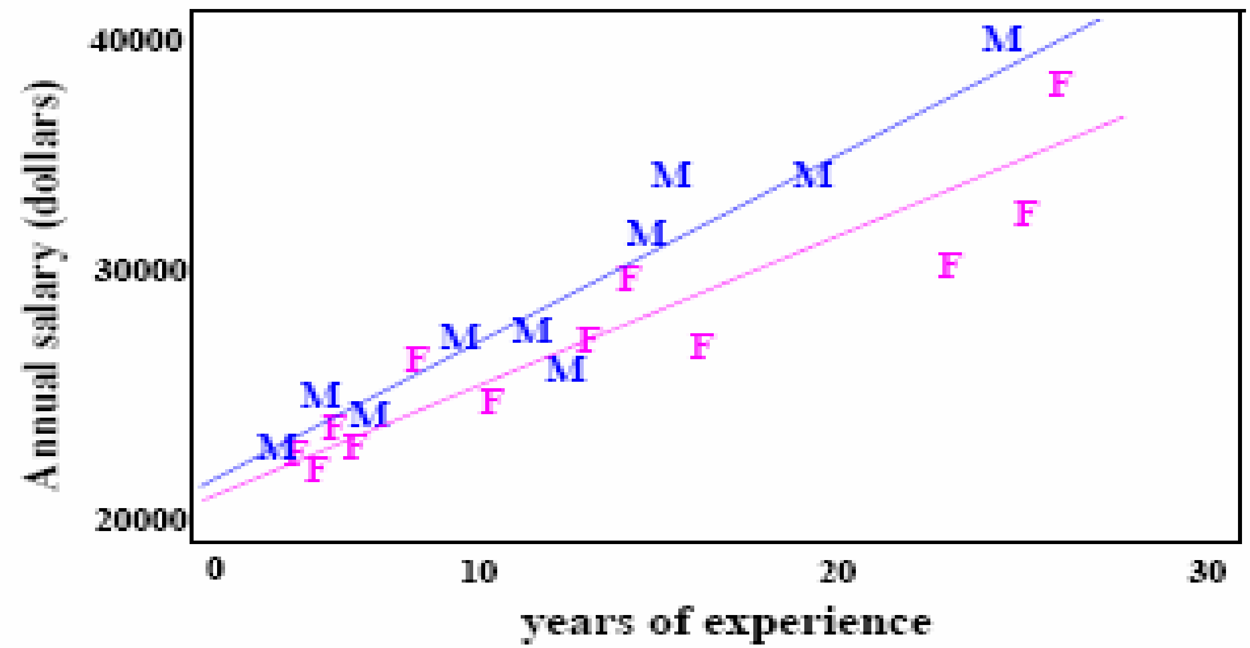
β_2

Slopes: β_1
Intercepts:
- Males: $\beta_0 + \beta_2$
- Females: β_0

- For the subpopulation of teachers at any particular years of experience, the mean salary for males is β_2 more than that for females.

Model with Interactions

$\mu(Y|X_1,X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(X_1 X_2)$
for all females: $\mu(Y|X_1,X_2=0) = \beta_0 + \beta_1 X_1$
for all males: $\mu(Y|X_1,X_2=1) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1$



Slopes:
- Males: $\beta_0 + \beta_2$
- Females: β_0

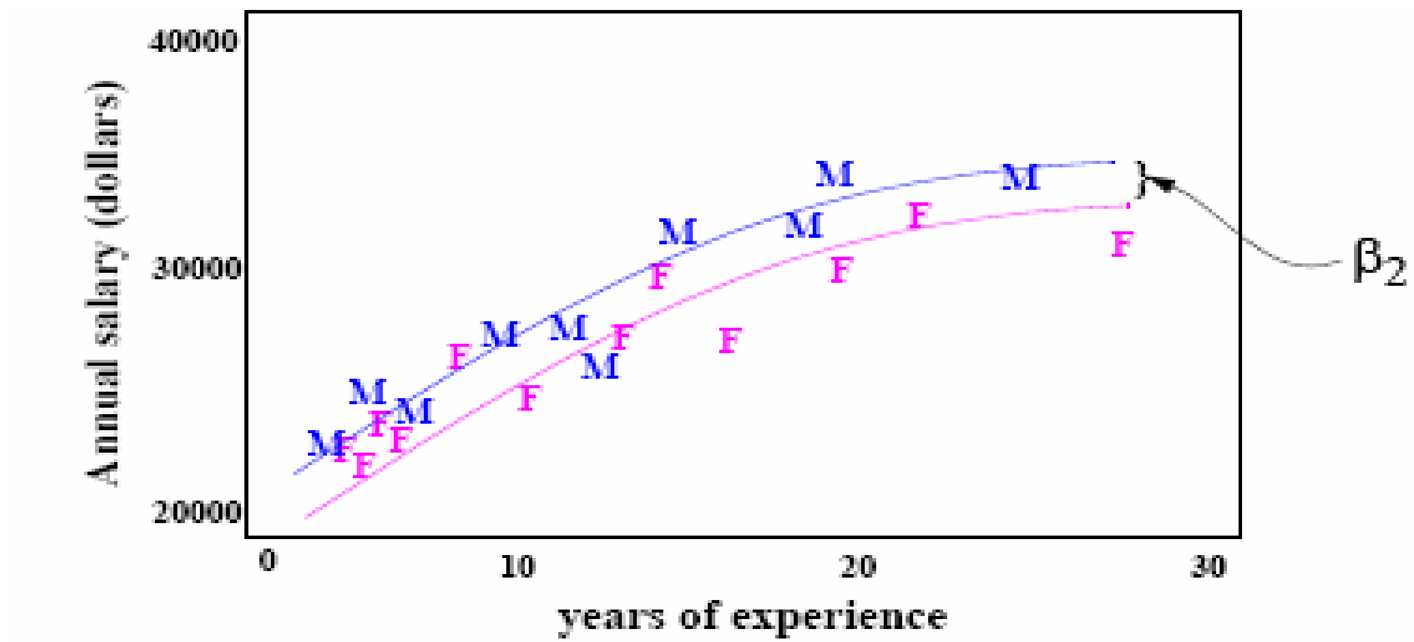
Intercepts:
- Males: $\beta_1 + \beta_3$
- Females: β_1

- The mean salary for inexperienced males (X_1=0) is β_2 (dollars) more than the mean salary for inexerienced females.
- The rate of increase in salary with increasing experience is β_3 (dollars) more for males than for females.

Model with curvilinear effects:

- Modelling curvature, parallel quadratic curves:

$$\mu(Y|X_1,X_2=1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2$$



- Modelling curvature, parallel quadratic curves:

$\mu(\text{salary}|..) = \beta_0 + \beta_1 \text{exper} + \beta_2 \text{Gender} + \beta_3 \text{exper}^2$

Notes about indicator variables

- A t-test for $H_0 : \beta_0=0$ in the regression of Y on a single indicator variable I_B, $\mu(Y|I_B) = \beta_0 + \beta_2 I_B$ is the 2-sample (difference of means) t-test
- Regression when all explanatory variables are categorical is "analysis of variance".
- Regression with categorical variables and one numerical X is often called "analysis of covariance".
- These terms are used more in the medical sciences than social science.
  - We'll just use the term "regression analysis" for all these variations.

Causation and Correlation

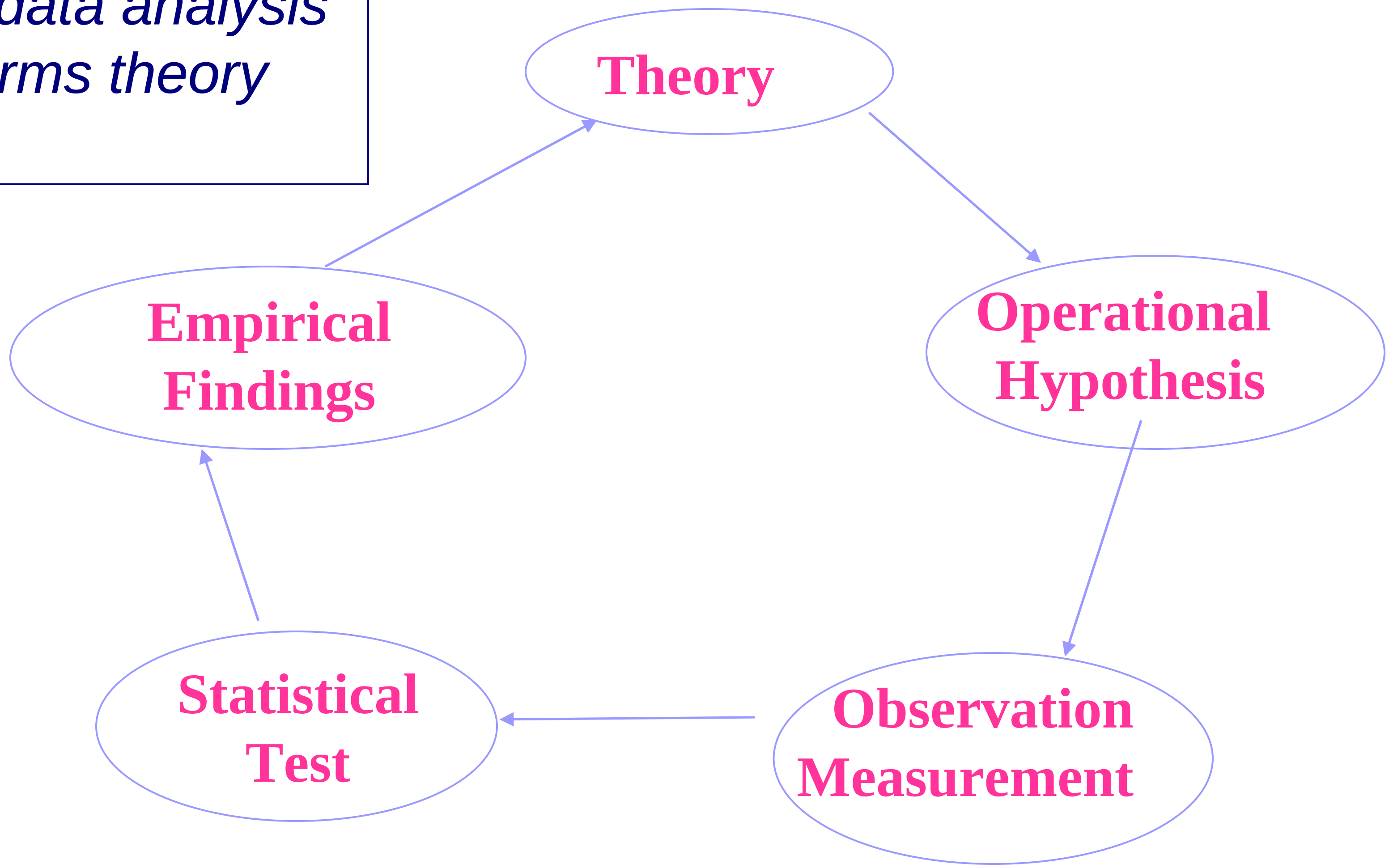
- Causal conclusions can be made from randomized experiments
  - But not from observational studies
- One way around this problem is to start with a <u>model</u> of your phenomenon
  - Then you test the implications of the model
  - These observations can <u>disprove</u> the model's hypotheses
    - But they cannot prove these hypotheses correct; they merely fail to reject the null

Models and Tests

- A **model** is an underlying theory about how the world works
  - Assumptions
  - Key players
  - Strategic interactions
  - Outcome set
- Models can be qualitative, quantitative, formal, experimental, etc.
  - But <u>everyone</u> uses models of some sort in their research
- Derive Hypotheses
  - E.g., as per capita GDP increases, countries become more democratic
- Test Hypotheses
  - Collect Data
    - Outcome and key explanatory variables
  - Identify the appropriate functional form
  - Apply the appropriate estimation procedures
  - Interpret the results

The traditional *scientific* approach

Virtuous cycle of theory informing data analysis which informs theory building



Example of a scientific approach

female education reduces childbearing

Women with higher education should have fewer children than those with less education

Using Ghana data? Women 15-49? Married or all women? How to measure education?

$CB_i = b_0 + b_1 * educ_i + resid_i$

Is b_1 significant? Positive, negative? Magnitude?

Strategies and Graphical Tools

Define the question of Interest
a) Specify theory
b) Hypothesis to be tested

Review Study Design
assumptions, logic, data availability, correct errors

1 **Explore the Data**

Use graphical tools; consider transformation; fit a tentative model; check outliers

2 **Formulate Inferential Model**
Derived from theory

State hypotheses in terms of model parameters

3 **Check Model**:
a) Model fit
b) Examine residuals
c) See if terms can be eliminated

Model Not OK

Check for non-constant variance; assess outliers

4 **Interpret results**
using appropriate tools

Confidence intervals, tests, prediction intervals

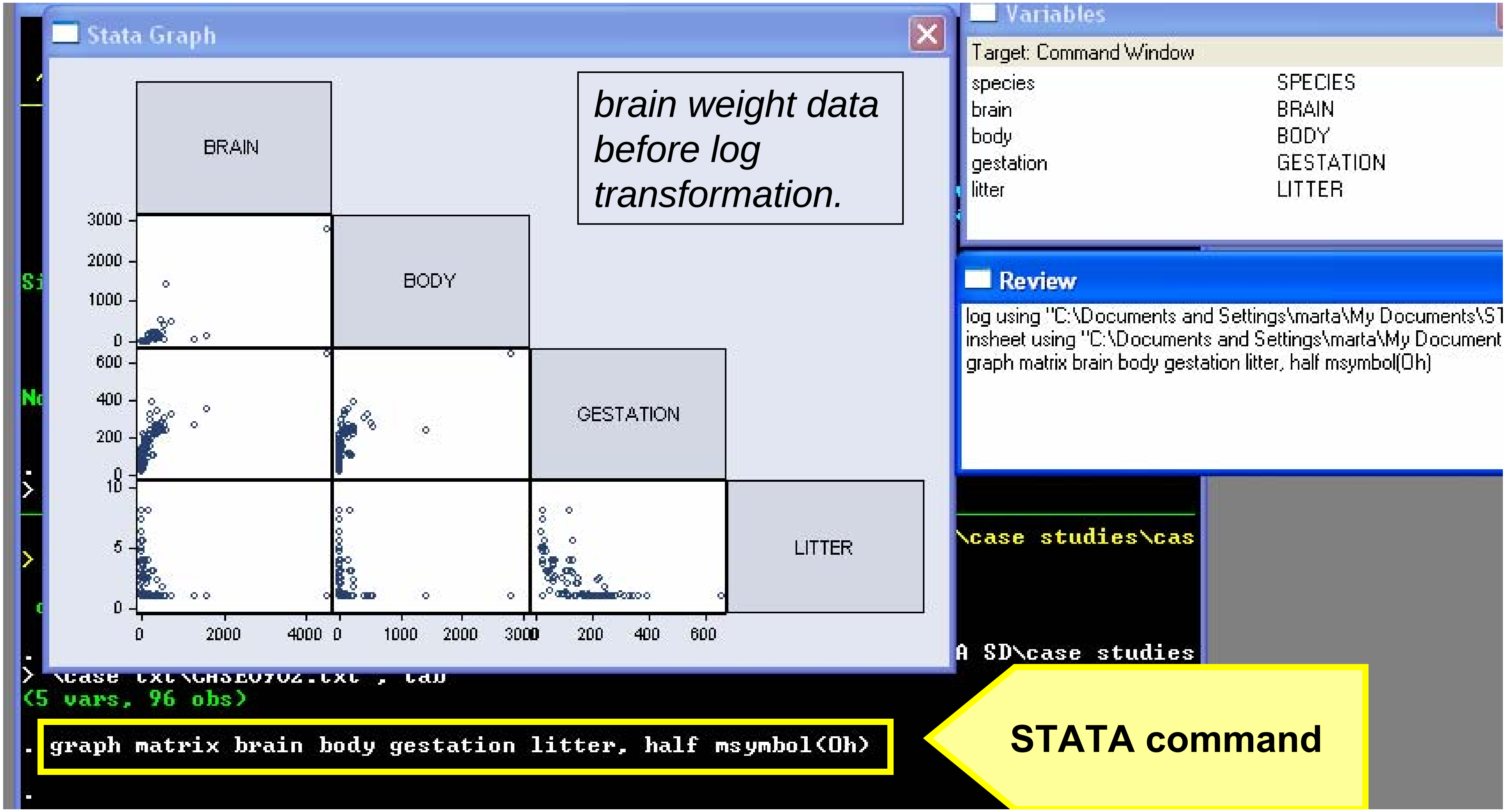
Presentation of results
Tables, graphs, text

Data Exploration

- Graphical tools for exploration and communication:
  - Matrix of scatterplots (9.5.1)
  - Coded scatterplot (9.5.2)
    - Different plotting codes for different categories
  - Jittered scatterplot (9.5.3)
  - Point identification
- Consider transformations
- Fit a tentative model
  - E.g., linear, quadratic, interaction terms, etc.
- Check outliers

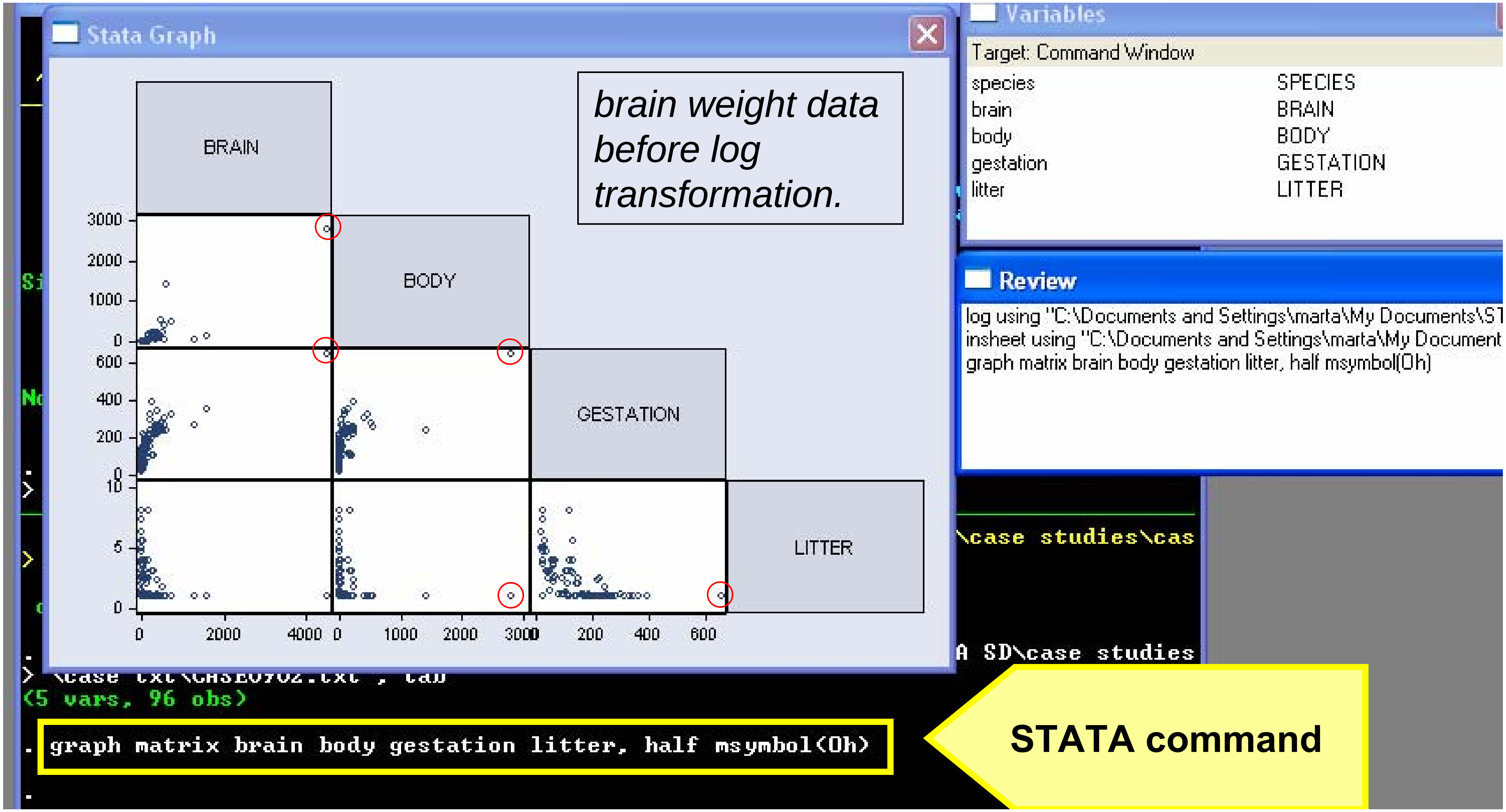
Scatter plots

Scatter plot matrices provide a compact display of the relationship between a number of variable pairs.

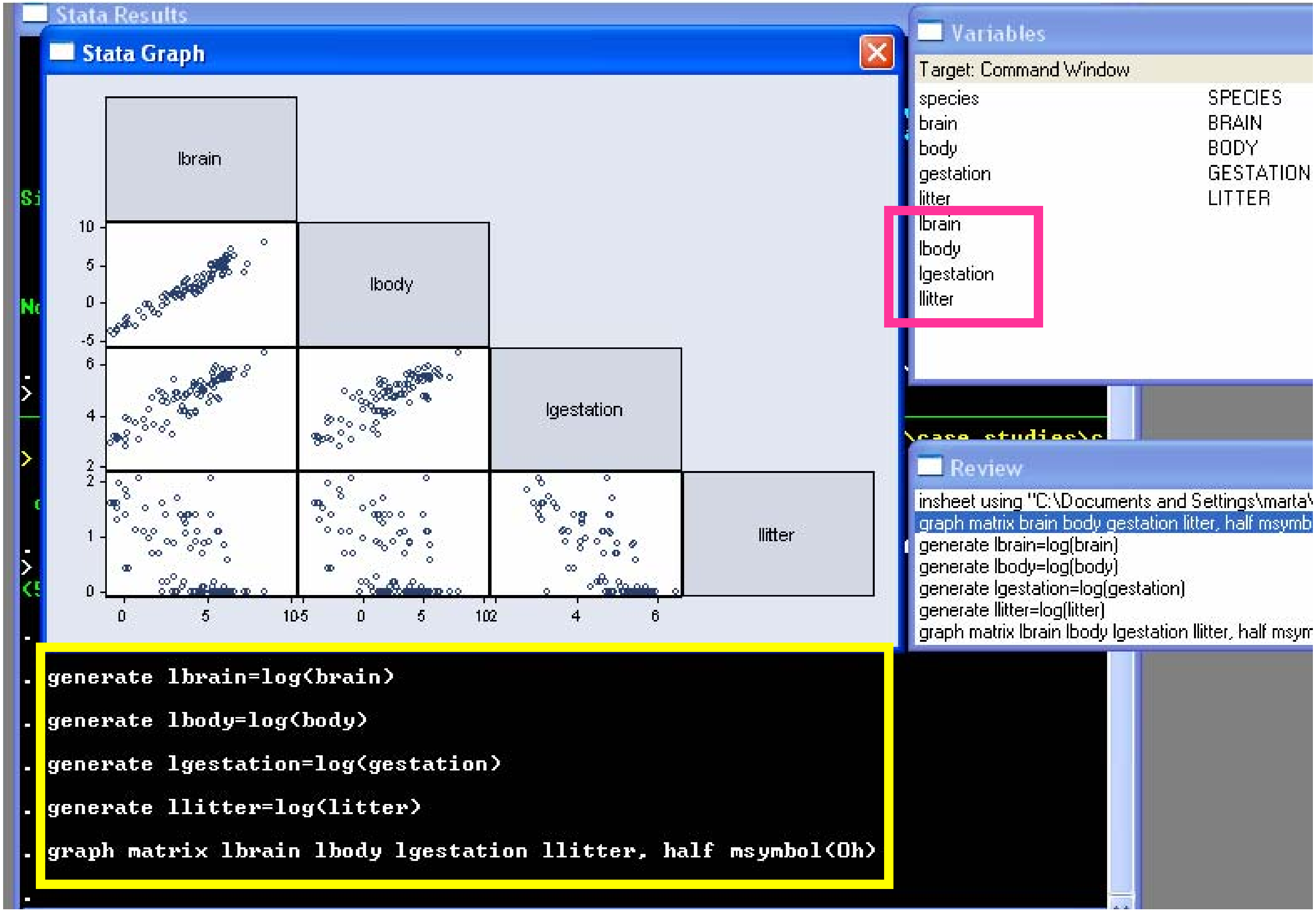


Scatter plots

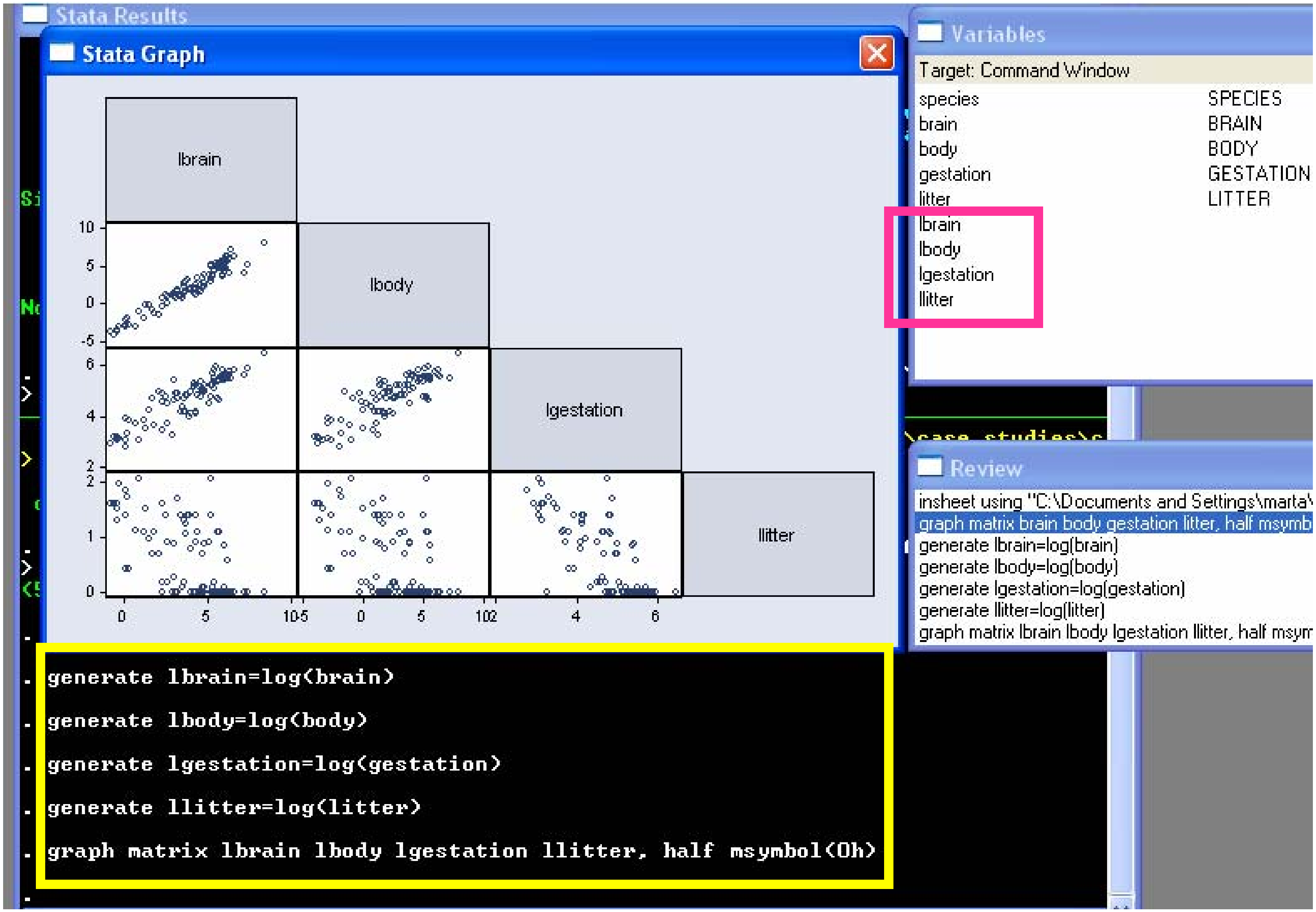
Scatter plot matrices can also indicate outliers



Scatterplot matrix for brain weight data after log transformation

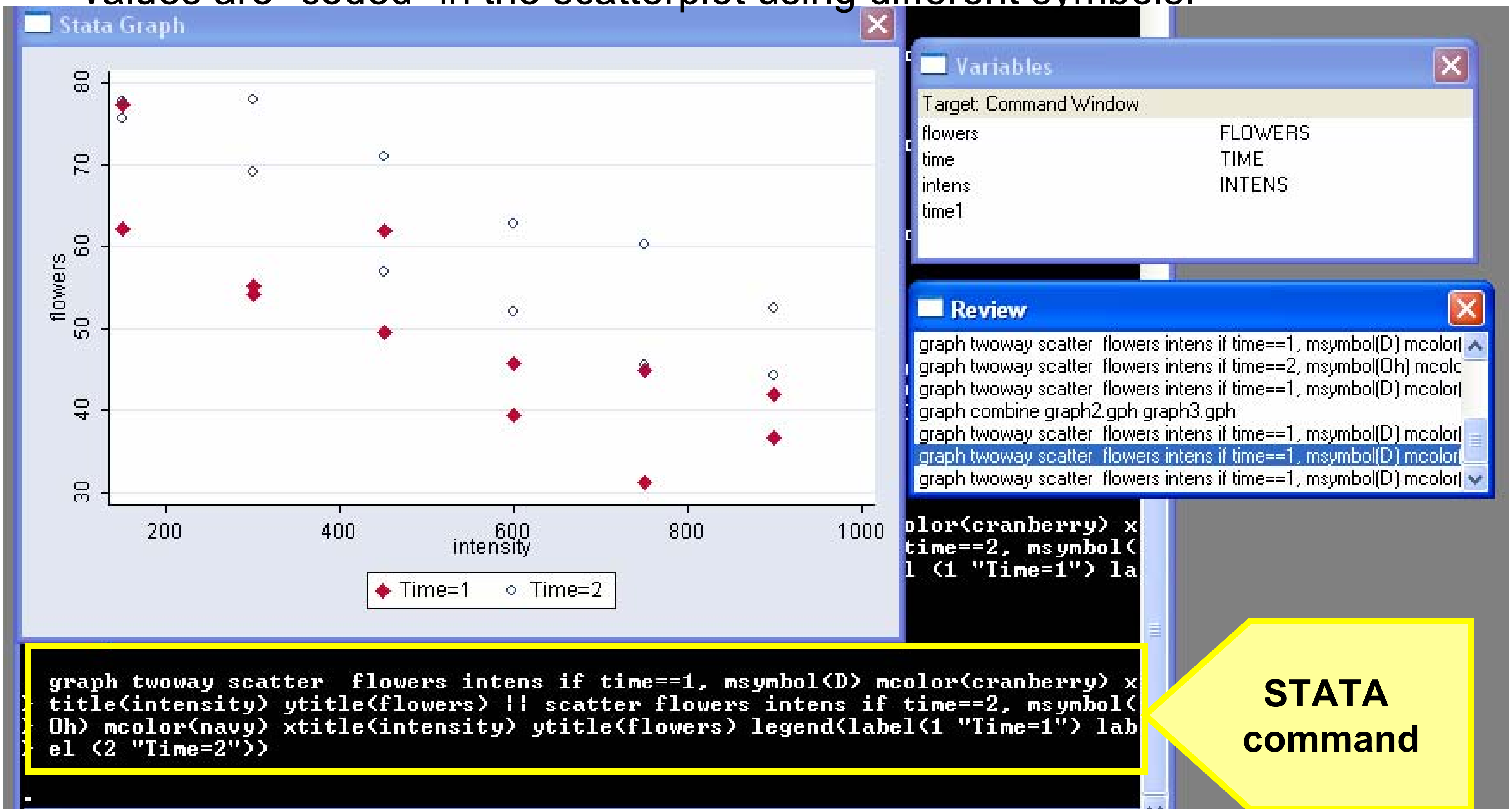


Notice: the outliers are now gone!



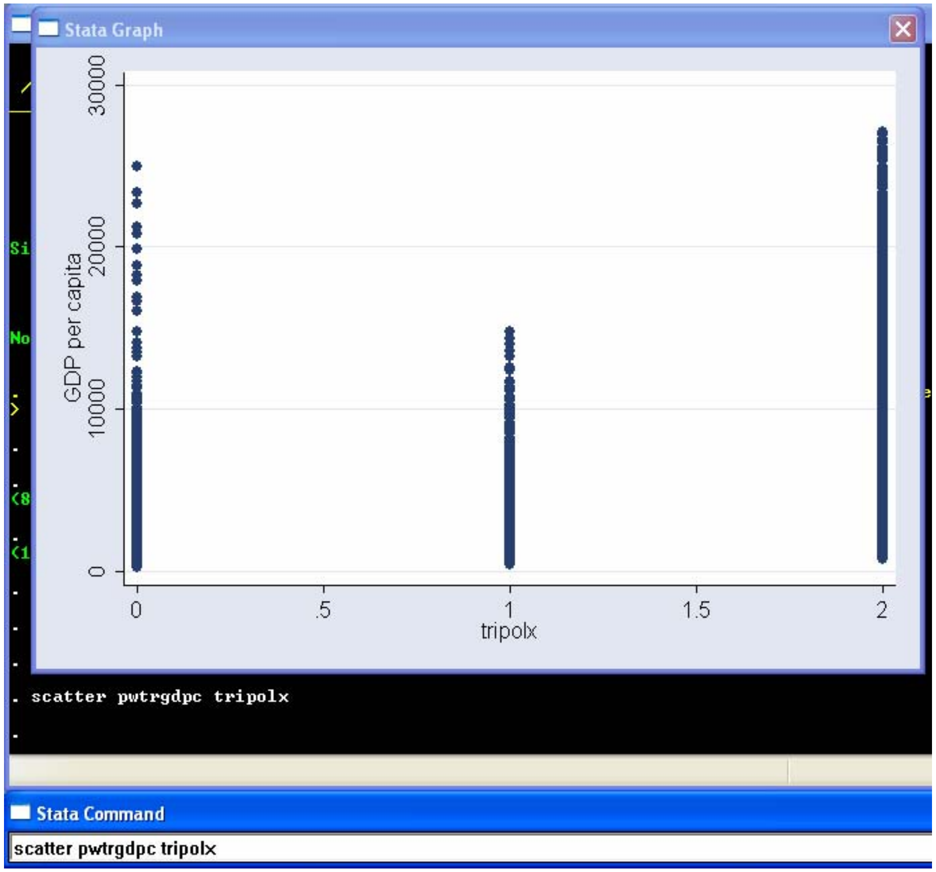
Coded Scatter Plots

- Coded scatter plots are obtained by using different plotting codes for different categories.
- In this example, the variable time has two possible values (1,2). Such values are "coded" in the scatterplot using different symbols.

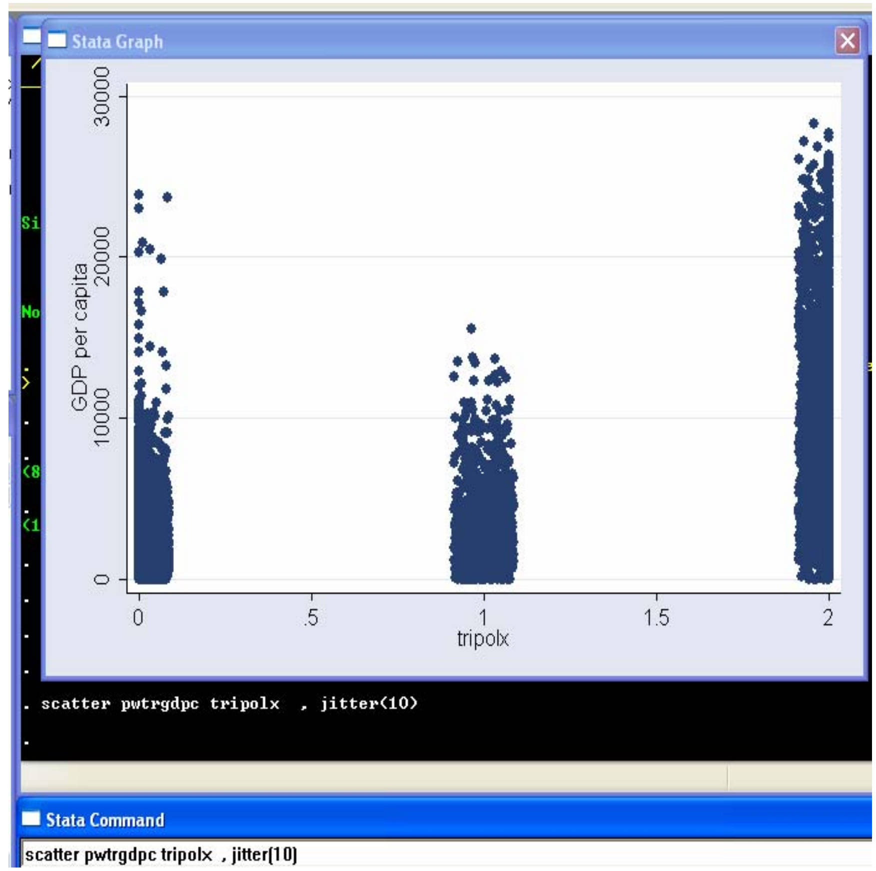


Jittering

Provides a clearer view of overlapping points.



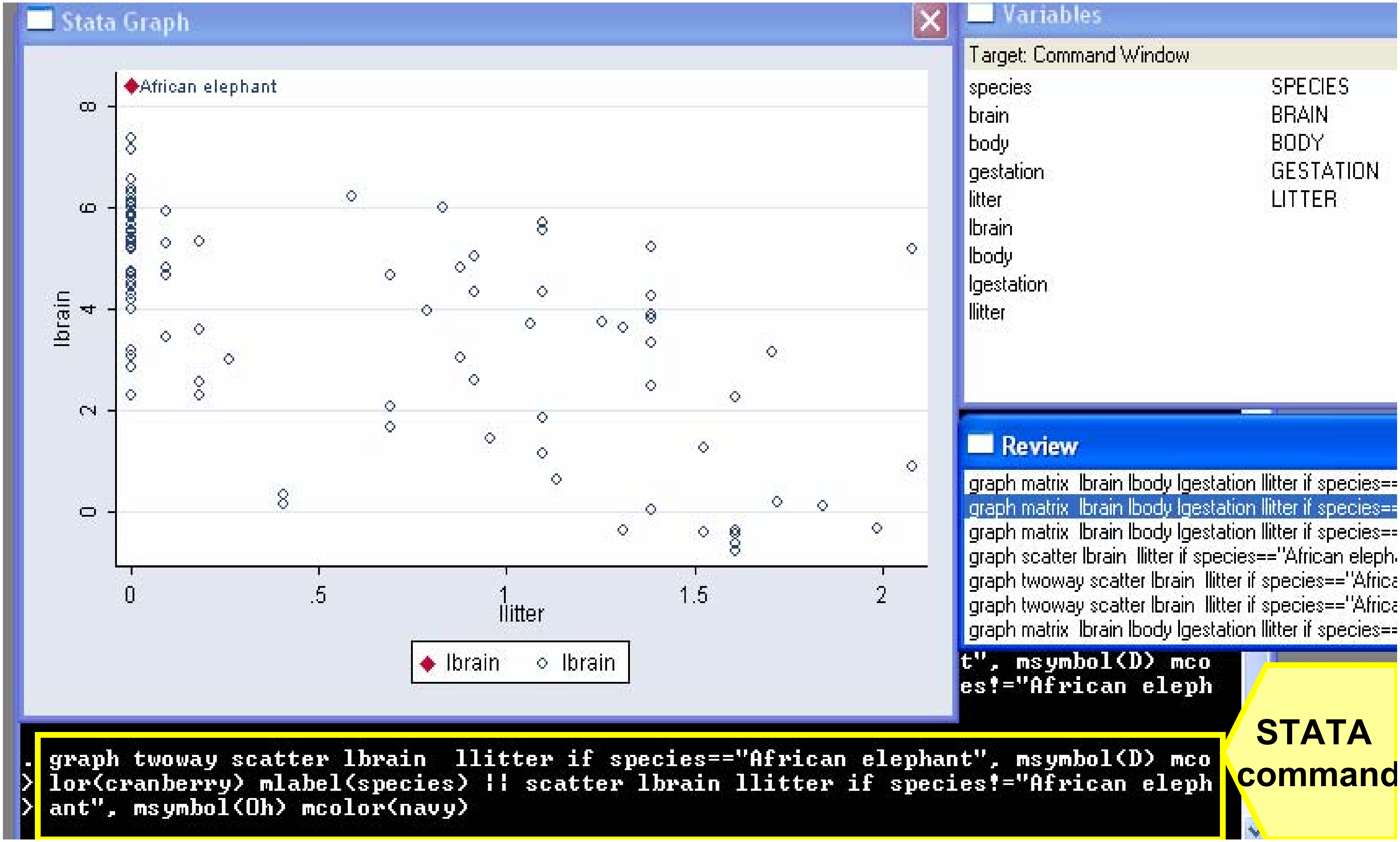
Un-jittered



Jittered

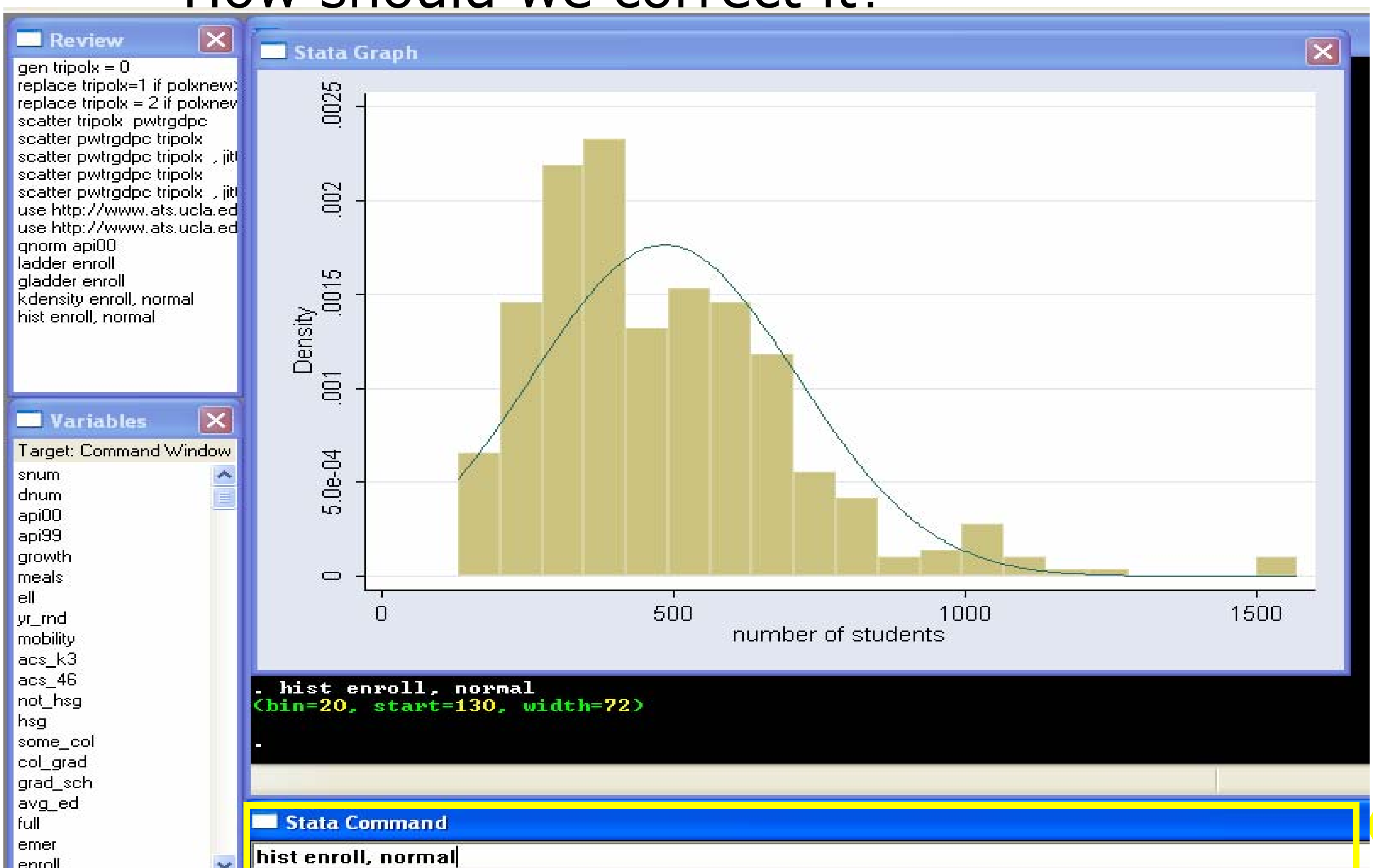
Point Identification

How to label points with STATA.



Transformations

This variable is clearly skewed –
How should we correct it?



STATA command

Transformations

Stata "ladder" command shows normality test for various transformations
Select the transformation with the lowest chi^2 statistic (this tests each distribution for normality)

```
. ladder enroll

Transformation         formula                   chi2(2)       P(chi2)
------------------------------------------------------------------
cubic                  enroll^3                       .         0.000
square                 enroll^2                       .         0.000
raw                    enroll                         .         0.000
square-root            sqrt(enroll)               20.56         0.000
log                    log(enroll)                 0.71         0.701
reciprocal root        1/sqrt(enroll)             23.33         0.000
reciprocal             1/enroll                   73.47         0.000
reciprocal square      1/(enroll^2)                   .         0.000
reciprocal cubic       1/(enroll^3)                   .         0.000
```

Transformations

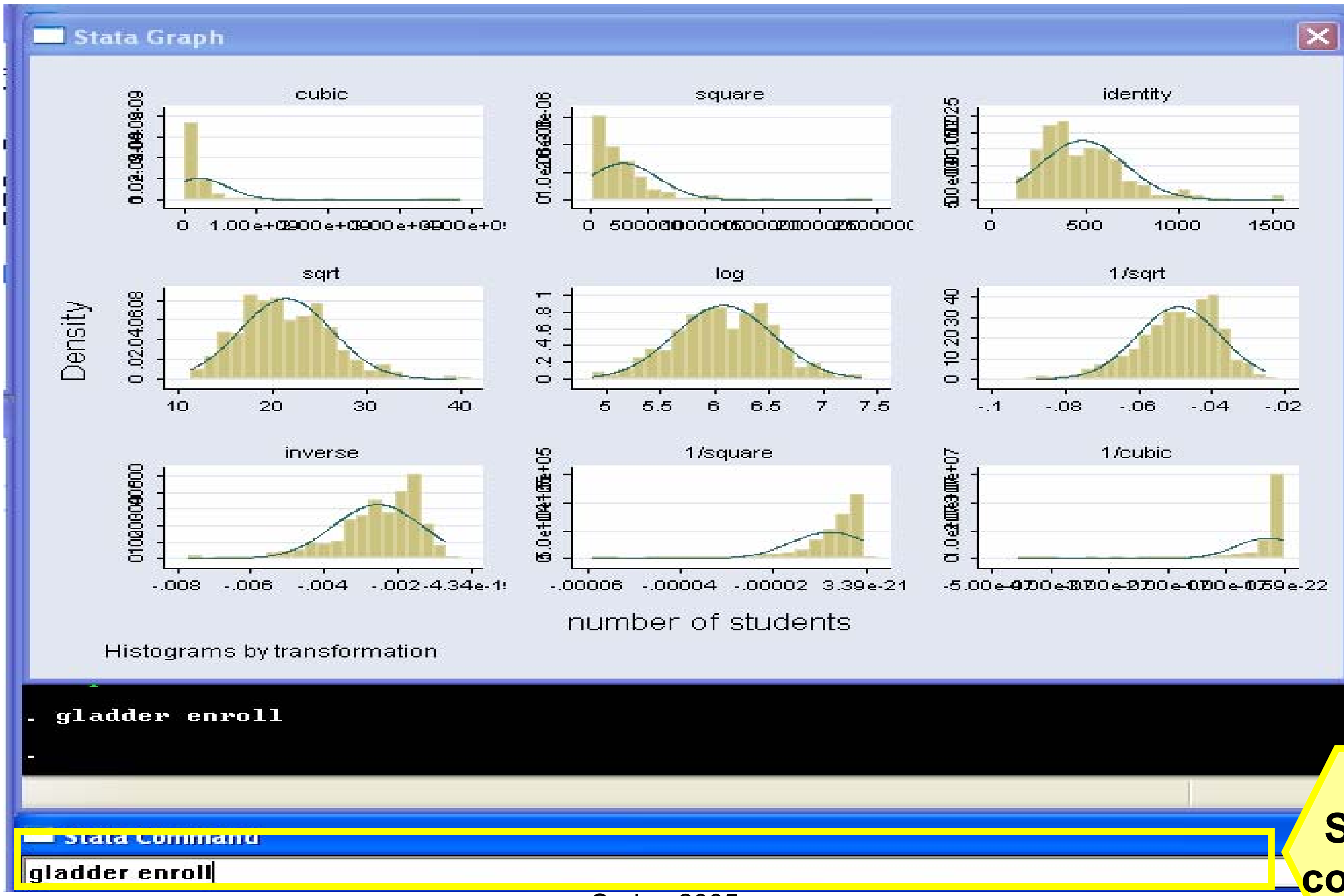
Stata "ladder" command shows normality test for various transformations
Select the transformation with the lowest chi^2 statistic (this tests each distribution for normality)

```
. ladder enroll
```

Transformation	formula	chi2(2)	P(chi2)
cubic	enroll^3	.	0.000
square	enroll^2	.	0.000
raw	enroll	.	0.000
square-root	sqrt(enroll)	20.56	0.000
log	log(enroll)	0.71	0.701
reciprocal root	1/sqrt(enroll)	23.33	0.000
reciprocal	1/enroll	73.47	0.000
reciprocal square	1/(enroll^2)	.	0.000
reciprocal cubic	1/(enroll^3)	.	0.000

Transformations

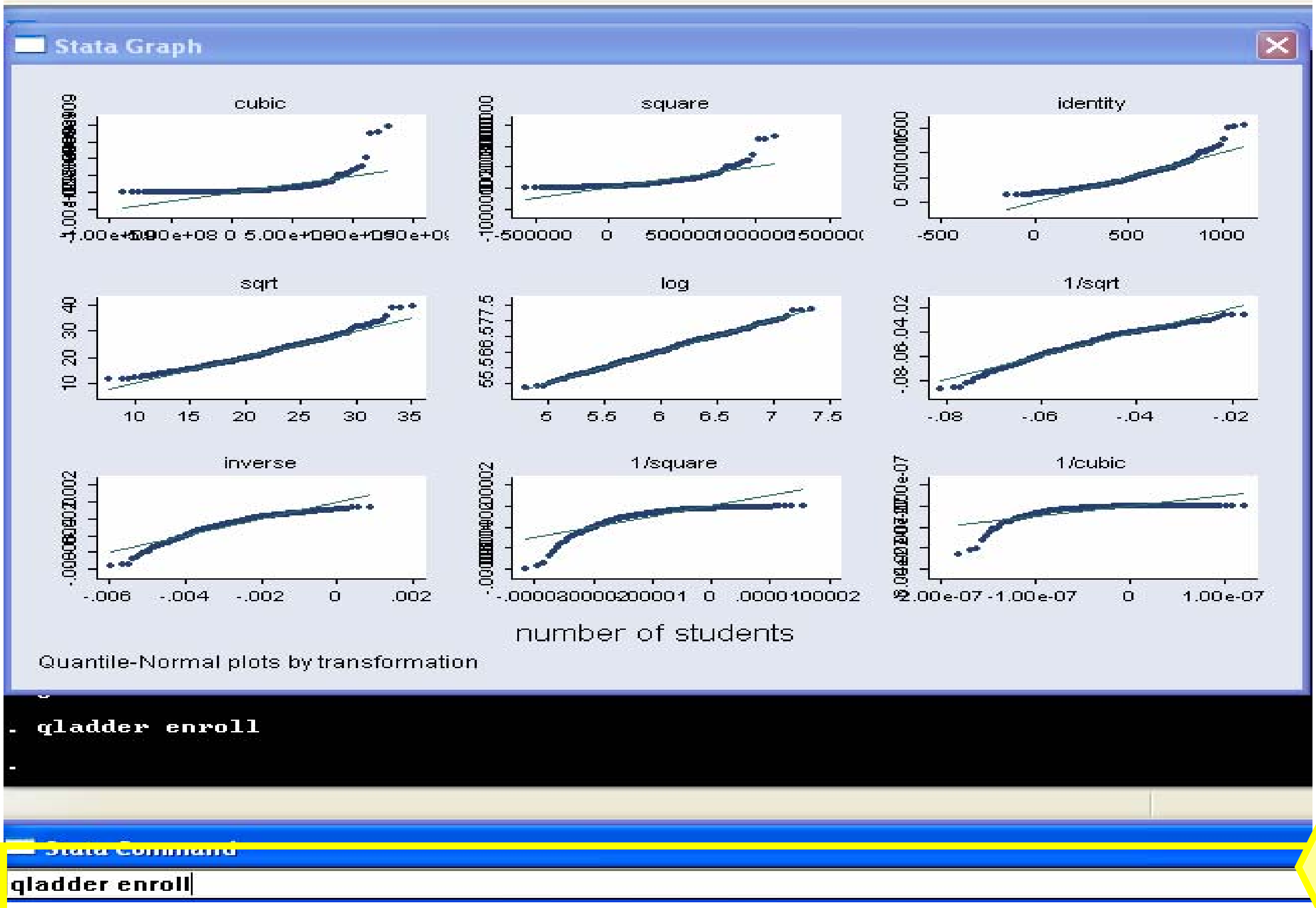
A graphical view of the different transformations using “gladder.”



STATA command

Transformations

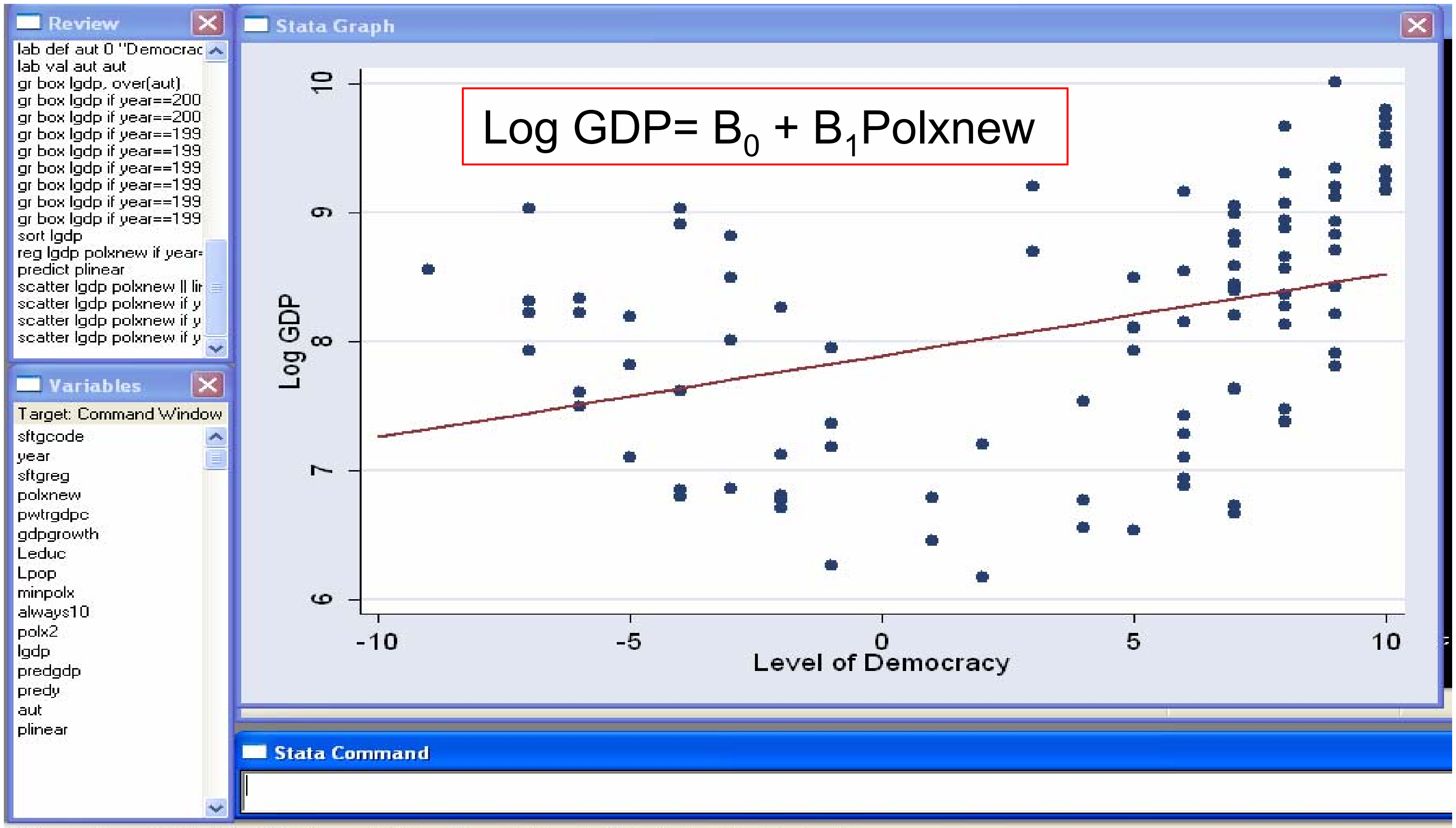
And yet another, using “qladder,” which gives a quantile-normal plot of each transformation



STATA command

Fit a Tentative Model

This models GDP and democracy, using only a linear term

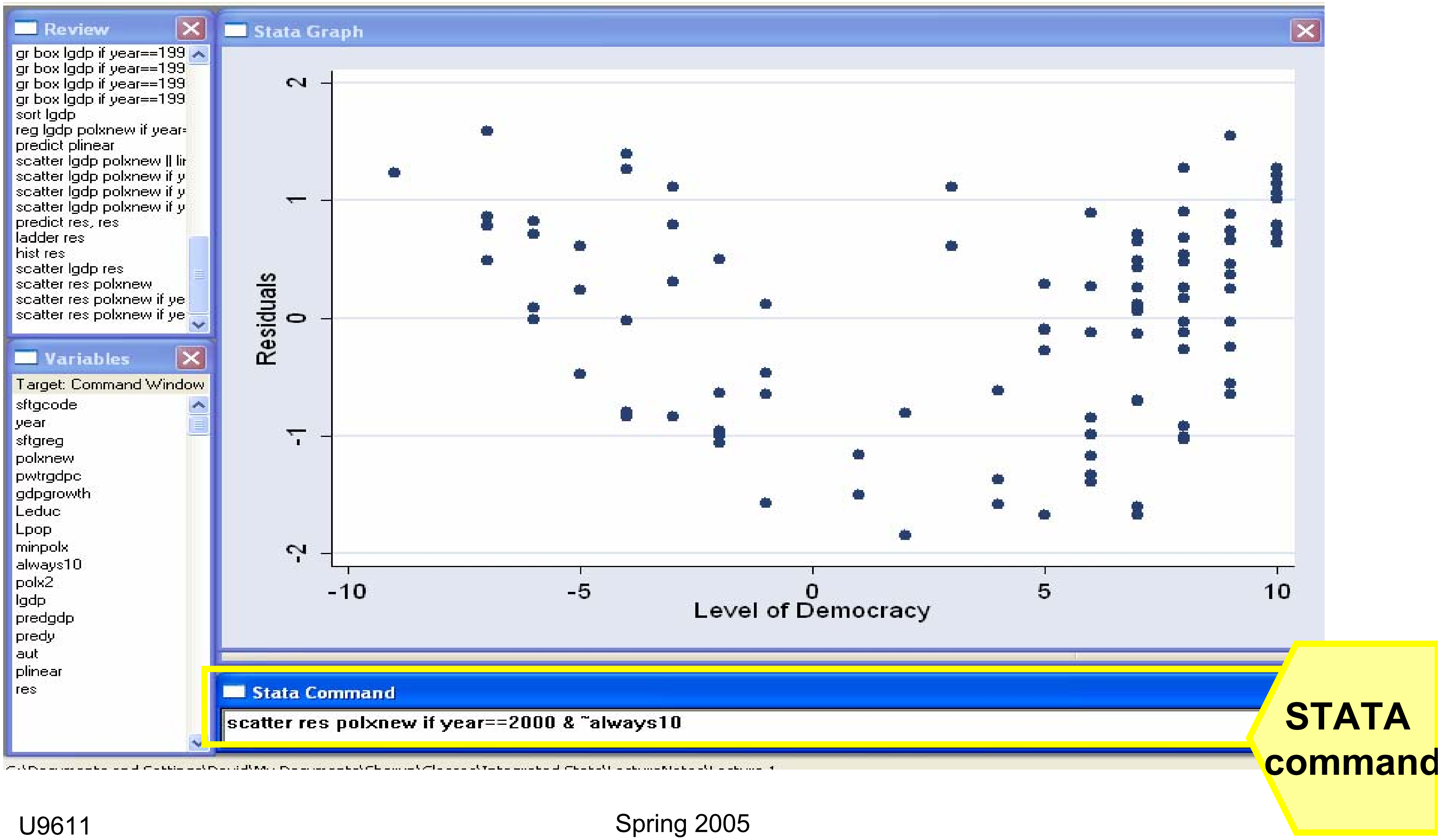


scatter lgdp polxnew if year==2000 & ~always10 || line plinear polxnew, sort legend(off) yti(Log GDP)

STATA command

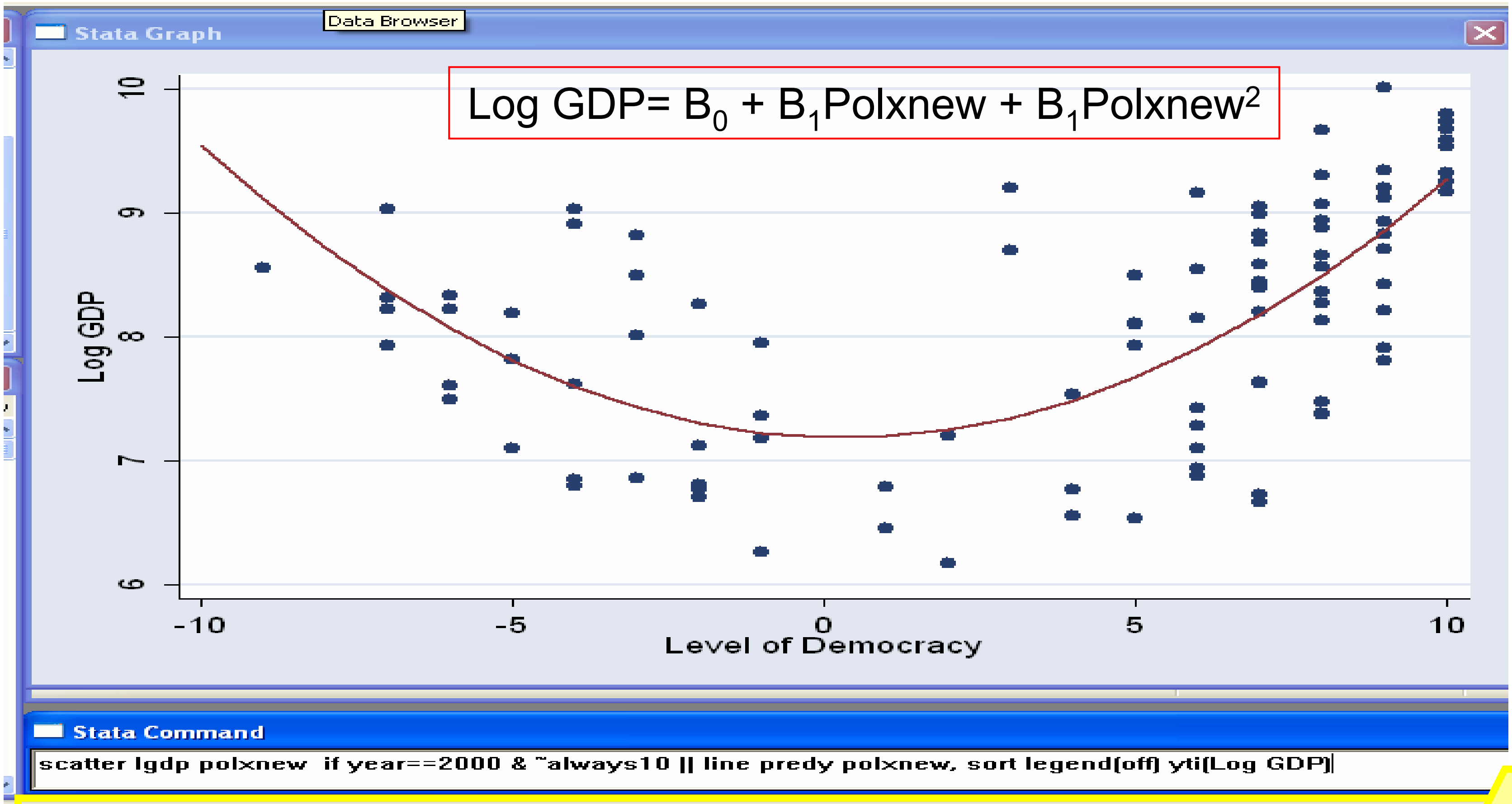
Fit a Tentative Model

The residuals from this regression are clearly U-shaped



Fit a Tentative Model

This models GDP and democracy, using a quadratic term as well

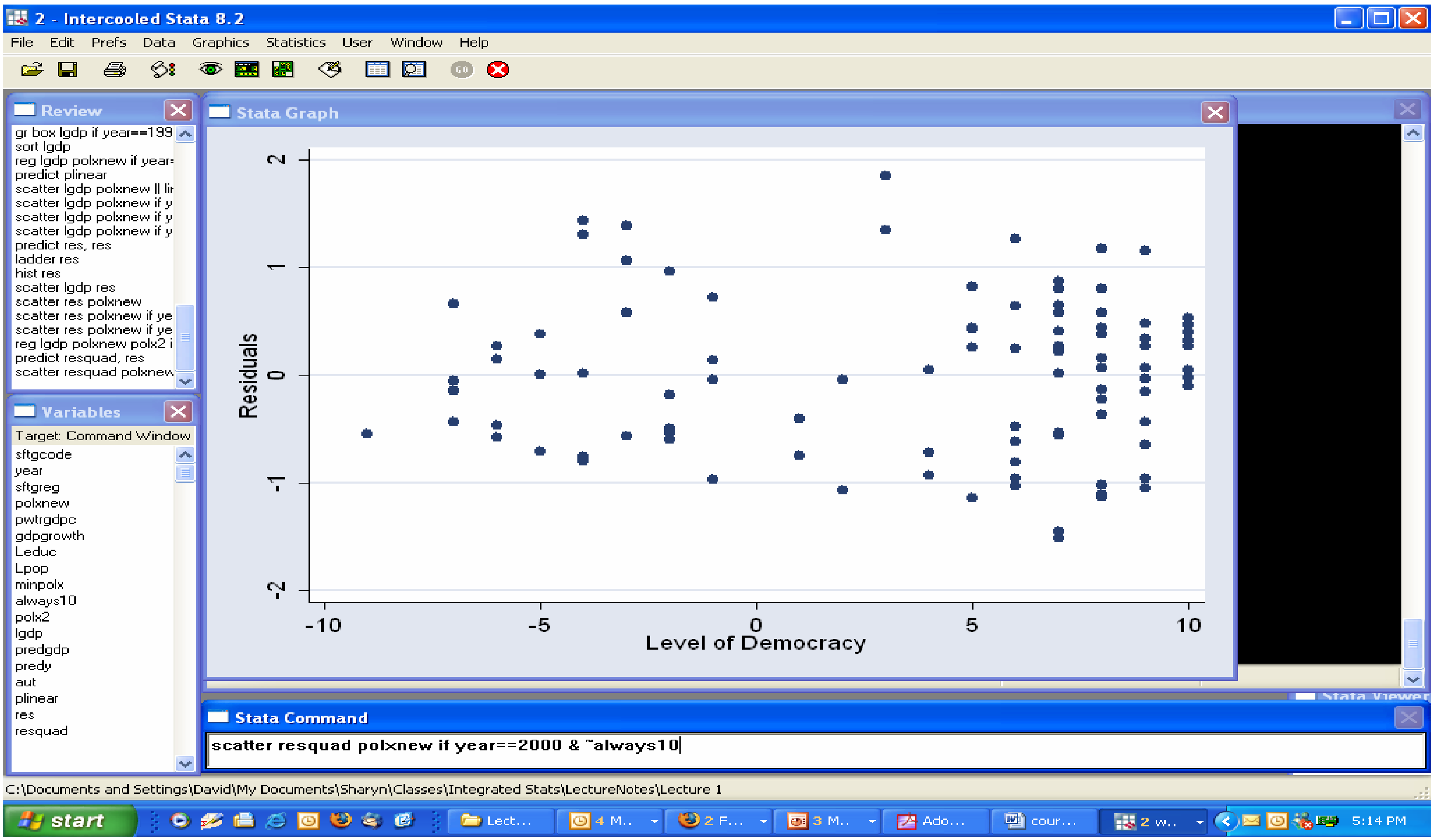


scatter lgdp polxnew if year==2000 & ~always10 || line predy polxnew, sort legend(off) yti(Log GDP)

STATA command

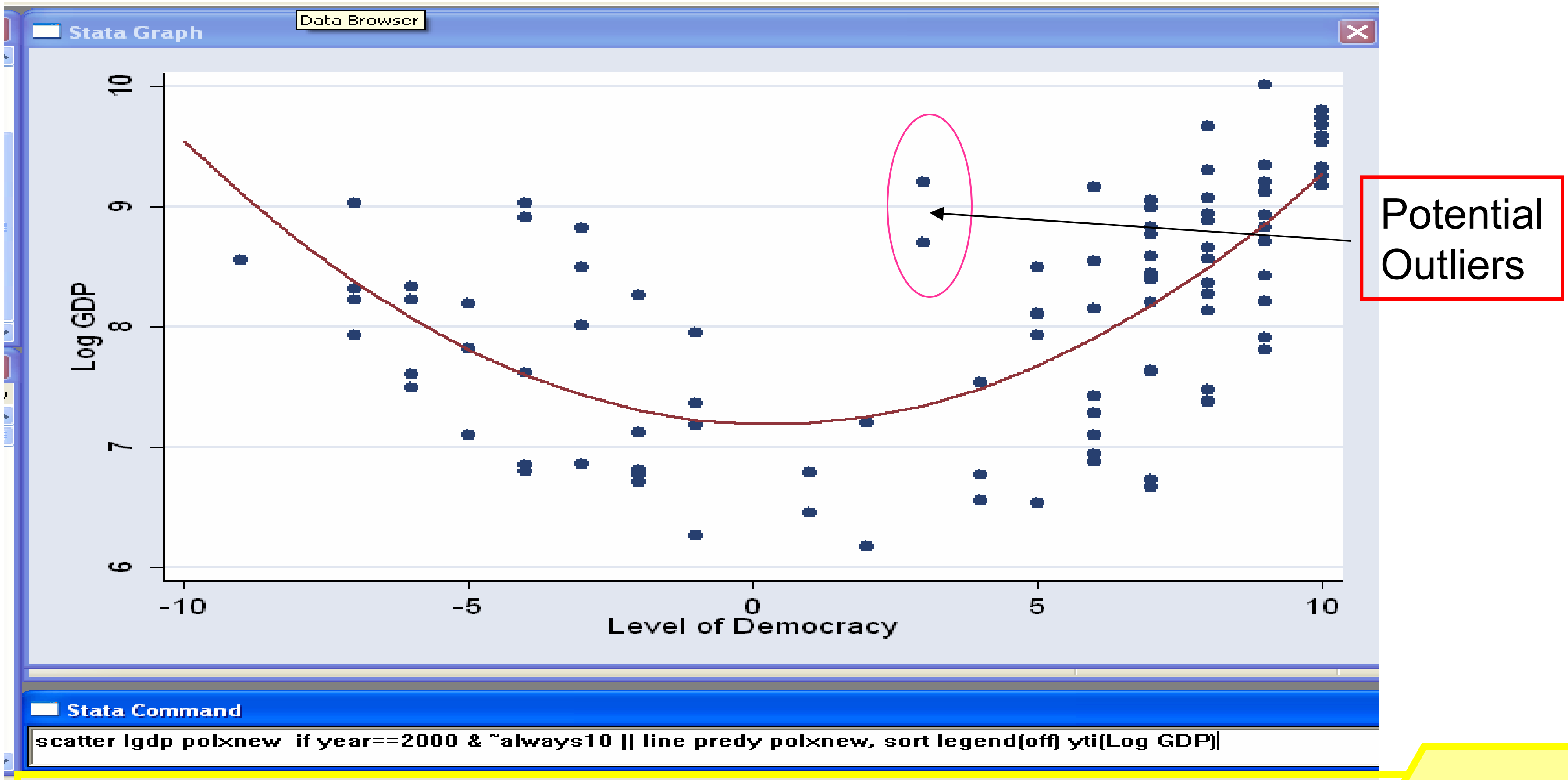
Fit a Tentative Model

Now the residuals look normally distributed



Check for Outliers

This models GDP and democracy, using a quadratic term

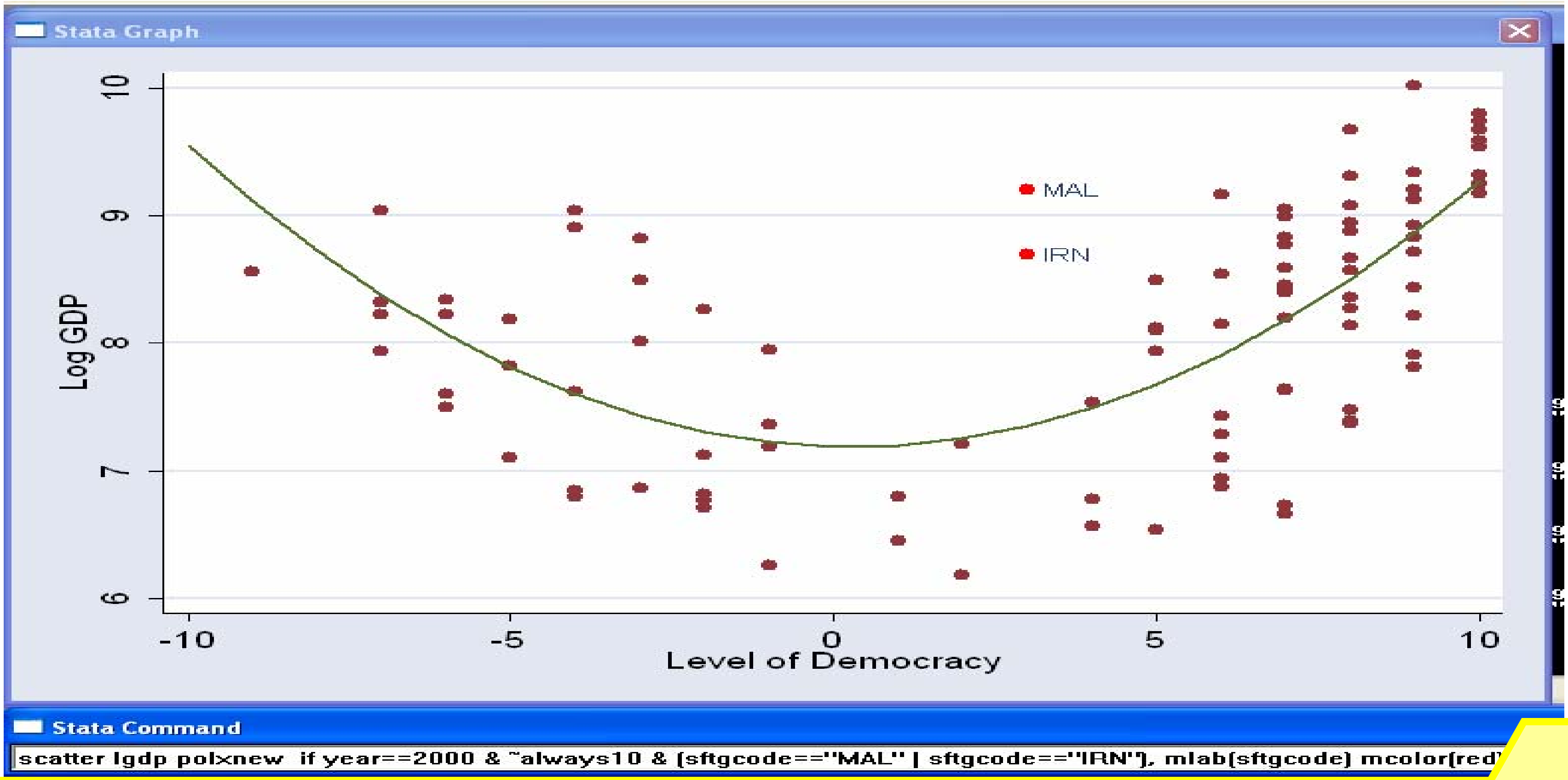



```
scatter lgdp polxnew  if year==2000 & ~always10 || line predy polxnew, sort
legend(off) yti(Log GDP)
```

STATA command

Check for Outliers

Identify outliers: Malawi and Iran




```
scatter lgdp polxnew  if year==2000 & ~always10 & (sftgcode=="MAL" | sftgcode=="IRN"),
mlab(sftgcode) mcolor(red) || scatter lgdp polxnew  if year==2000 & ~always10 & (sftgcode!="MAL" &
sftgcode!="IRN") || line predy polxnew, sort legend(off) yti(Log GDP)
```

STATA command

Check for Outliers

```
. reg lgdp polxnew polx2 if year==2000 & ~always10

      Source |       SS       df       MS              Number of obs =      97
-------------+------------------------------           F(  2,    94) =   34.84
       Model |  36.8897269     2  18.4448635           Prob > F      =  0.0000
    Residual |  49.7683329    94   .52945035           R-squared     =  0.4257
-------------+------------------------------           Adj R-squared =  0.4135
       Total |  86.6580598    96  .902688123           Root MSE      =  .72763

------------------------------------------------------------------------------
        lgdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     polxnew |  -.0138071   .0173811    -0.79   0.429    -.0483177    .0207035
       polx2 |    .022208   .0032487     6.84   0.000     .0157575    .0286584
       _cons |   7.191465   .1353228    53.14   0.000     6.922778    7.460152
------------------------------------------------------------------------------
```

Try analysis without the outliers; same results.

```
. reg lgdp polxnew polx2 if year==2000 & ~always10 & (sftgcode!="MAL" & sftgcode!="IRN")

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F(  2,    92) =   42.67
       Model |  40.9677226     2  20.4838613           Prob > F      =  0.0000
    Residual |   44.164877    92  .480053011           R-squared     =  0.4812
-------------+------------------------------           Adj R-squared =  0.4699
       Total |  85.1325996    94  .905665953           Root MSE      =  .69286

------------------------------------------------------------------------------
        lgdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     polxnew |  -.0209735   .0166859    -1.26   0.212    -.0541131    .0121661
       polx2 |   .0244657   .0031649     7.73   0.000       .01818    .0307514
       _cons |   7.082237   .1328515    53.31   0.000     6.818383    7.346092
------------------------------------------------------------------------------
```

So leave in model;

See Display 3.6 for other strategies.

EXAMPLE: Rainfall and Corn Yield

(Exercise: 9.15, page 261)

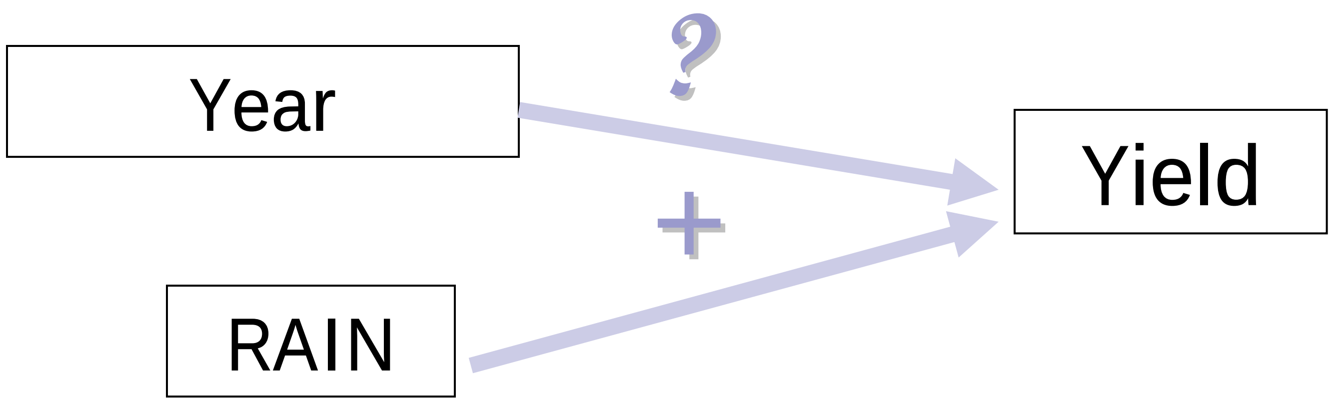
Dependent variable (Y): Yield
Explanatory variables (Xs):

- Rainfall
- Year

- Linear regression (scatterplot with linear regression line)
- Quadratic model (scatter plot with quadratic regression curve)
- Conditional scatter plots for yield vs. rainfall (selecting different years)
- Regression model with quadratic functions and interaction terms

Model of Rainfall and Corn Yield

- Let's say that we collected data on corn yields from various farms.
  - Varying amounts of rainfall could affect yield.
  - But this relation may change over time.
- The causal model would then look like this:



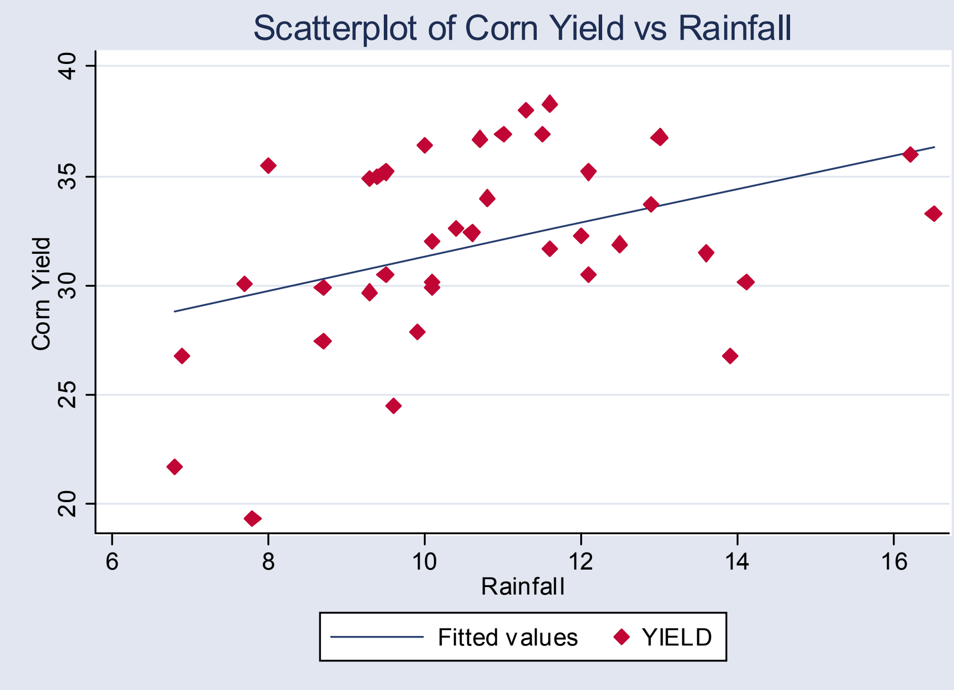
Scatterplot

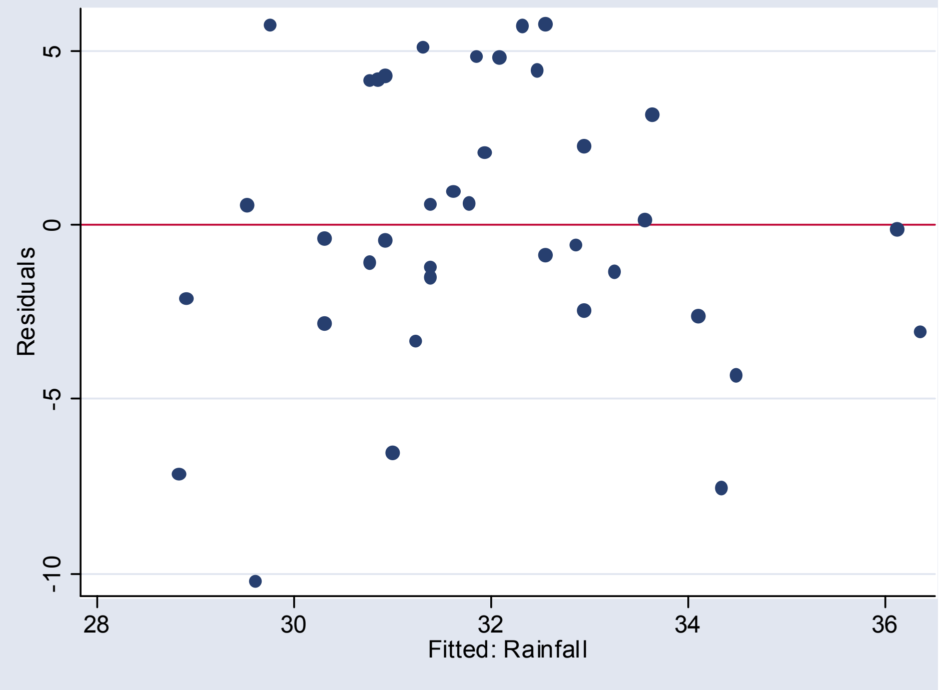
Initial scatterplot of yield vs rainfall, and residual plot from simple linear regression fit.

```
reg yield rainfall

graph twoway lfit yield rainfall || scatter  yield rainfall, msymbol(D)
mcolor(cranberry) ytitle("Corn yield") xtitle("Rainfall")
title("Scatterplot of Corn Yield vs Rainfall")

rvfplot, yline(0) xtitle("Fitted: Rainfall")
```

$$Yield=\beta_0+\beta_1 rainfall$$

STATA command

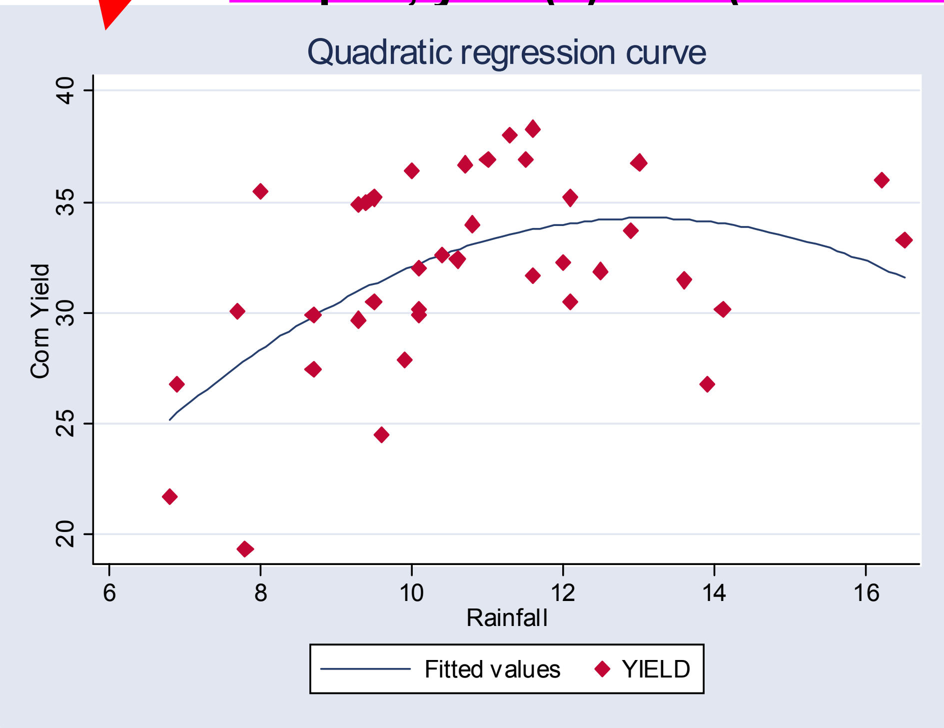


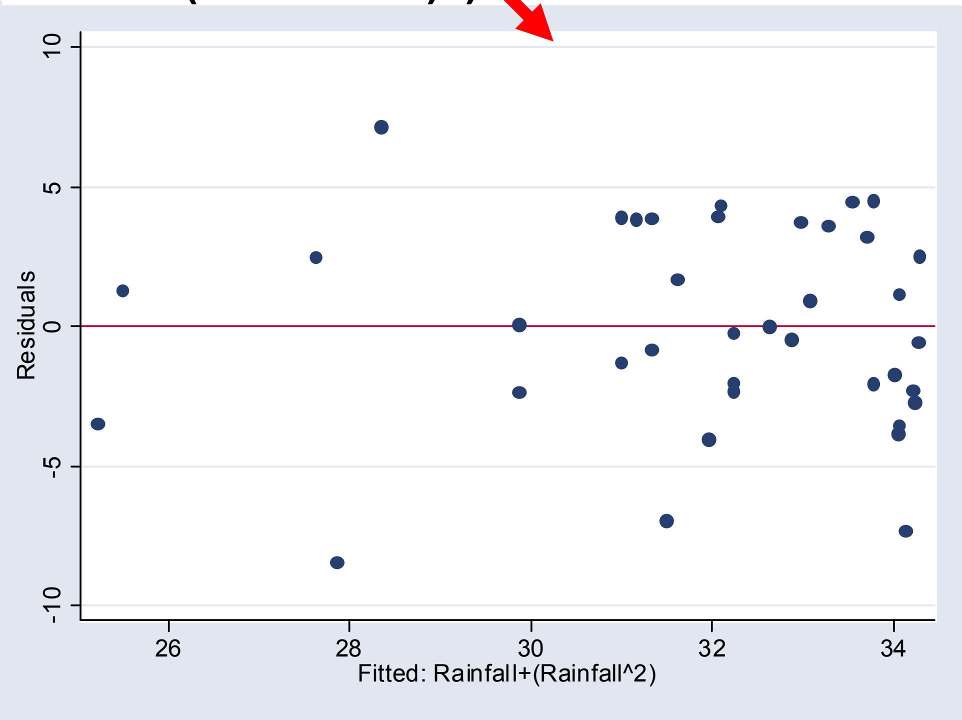


Quadratic fit: represents better the yield-trend

```
graph twoway qfit yield rainfall || scatter  yield rainfall, msymbol(D)
mcolor(cranberry) ytitle("Corn Yield") xtitle("Rainfall")
title("Quadratic regression curve")

gen rainfall2=rainfall^2

reg yield rainfall rainfall 2

rvfplot, yline(0) xtitle("Fitted: Rainfall+(Rainfall^2)")
```

$$Yield = \beta_0 + \beta_1 rainfall + \beta_2 rainfall^2$$





Quadratic fit: Residual plot vs time

Since data were collected over time we should check for time trend and serial correlation, by plotting residuals vs. time.

$Yield = \beta_0 + \beta_1 rainfall + \beta_2 rainfall^2$

```
. insheet using "C:\Documents and Settin
> xt\EX0915.txt", tab
(3 vars, 38 obs)

. gen rainfall2=rainfall^2

. gen rainyear=rainfall*year

. reg yield rainfall rainfall2

      Source |       SS       df       MS              Number of obs =      38
-------------+------------------------------           F(  2,    35) =    7.38
       Model |  209.021698     2  104.510849           Prob > F      =  0.0021
    Residual |  495.528887    35  14.1579682           R-squared     =  0.2967
-------------+------------------------------           Adj R-squared =  0.2565
       Total |  704.550584    37  19.0419077           Root MSE      =  3.7627

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rainfall |   6.004282    2.03895     2.94   0.006     1.864994    10.14357
   rainfall2 |  -.2293639    .088635    -2.59   0.014    -.4093025   -.0494252
       _cons |  -5.014659   11.44158    -0.44   0.664    -28.2423     18.21298
------------------------------------------------------------------------------

. predict e, resid

. label variable e "Residual for model (rain+rain^2)"

. graph twoway lfit e year || scatter e year
```

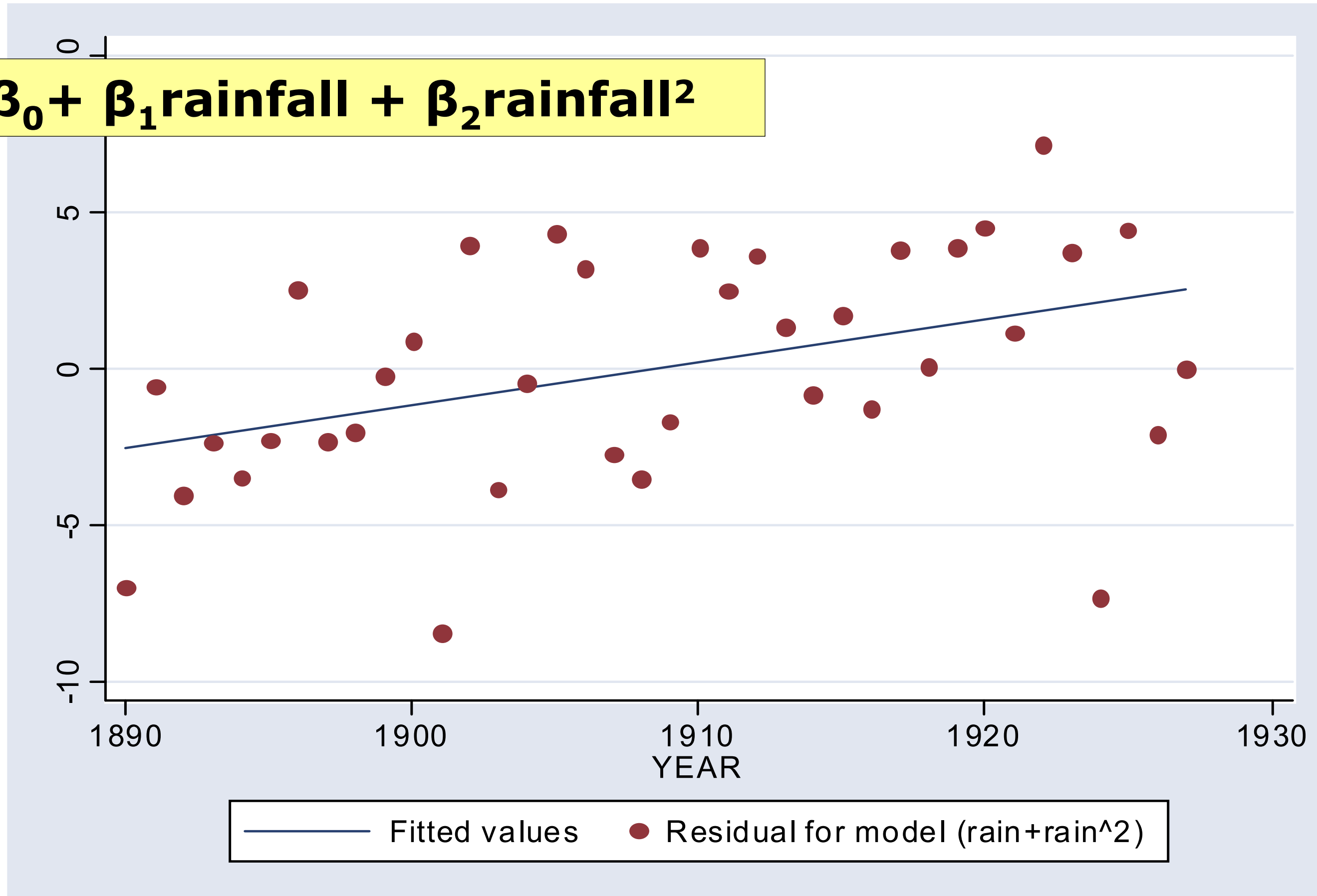
Target: Command Window

year	YEAR
yield	YIELD
rainfall	RAINFALL
rainfall2	
rainyear	
e	Residual for model (rain+rain^2)

1. **Run regression**
2. **Predict residuals**
3. **Graph scatterplot residuals vs. time**

Graph: Scatterplot residuals vs. year

Yield=β_0+ $\beta_1$rainfall + $\beta_2$rainfall^2



- There does appear to be a trend.
- There is no obvious serial correlation. (more in Ch. 15)
- Note: **Year** is not an explanatory variable in the regression model.

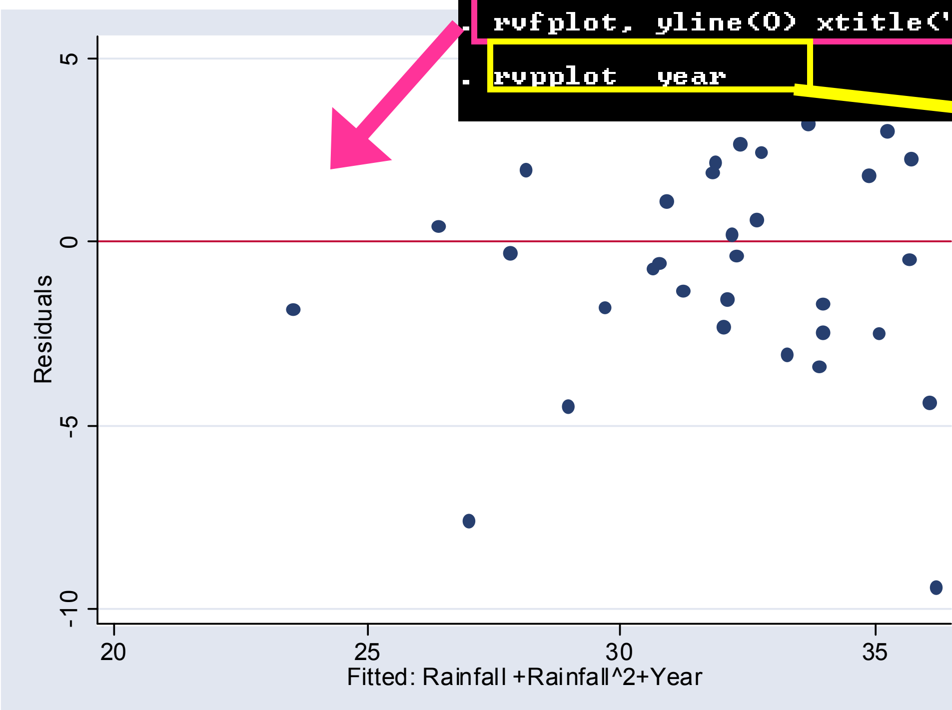
Adding time trend

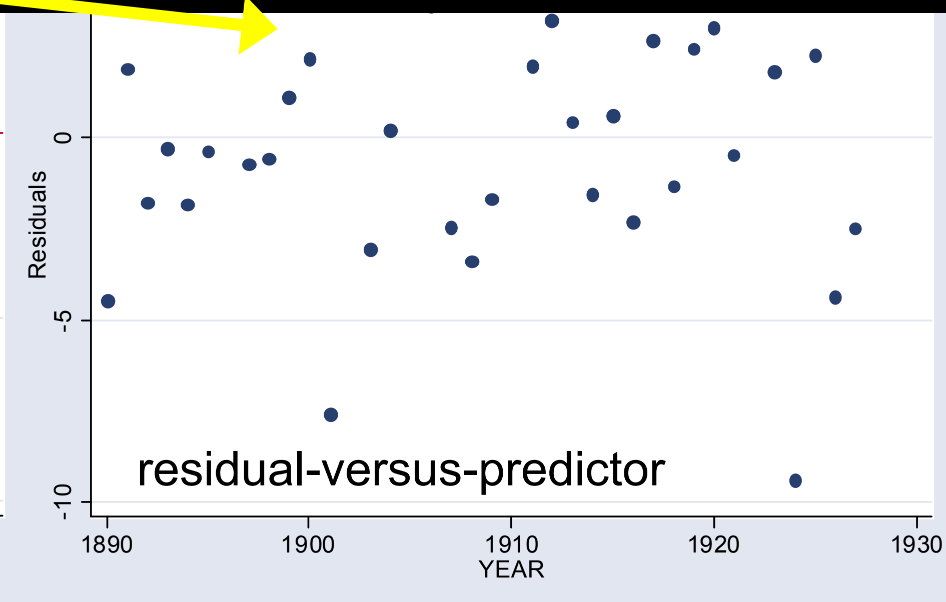
```
. reg yield  rainfall  rainfall2 year
```

Yield=β_0+ $\beta_1$rainfall + $\beta_2$rainfall^2+ β_3Year

Include Year in the regression model

```
    Residual |  410.993245    34  12.0880366           R-squared     =  0.4167
-------------+------------------------------           Adj R-squared =  0.3652
       Total |  704.550584    37  19.0419077           Root MSE      =  3.4768

                                                       Number of obs =      38
                                                       F(  3,    34) =    8.09
                                                       Prob > F      =  0.0003

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rainfall |    5.67038   1.888239     3.00   0.005     1.833016    9.507744
   rainfall2 |   -.215498   .0820675    -2.63   0.013    -.3822791   -.0487169
        year |   .1363414   .0515568     2.64   0.012     .0315654    .2411174
       _cons |  -263.3032   98.24096    -2.68   0.011    -462.9529   -63.65359
------------------------------------------------------------------------------

. rvfplot, yline(0) xtitle("Fitted: Rainfall +Rainfall^2+Year")

. rvpplot  year
```



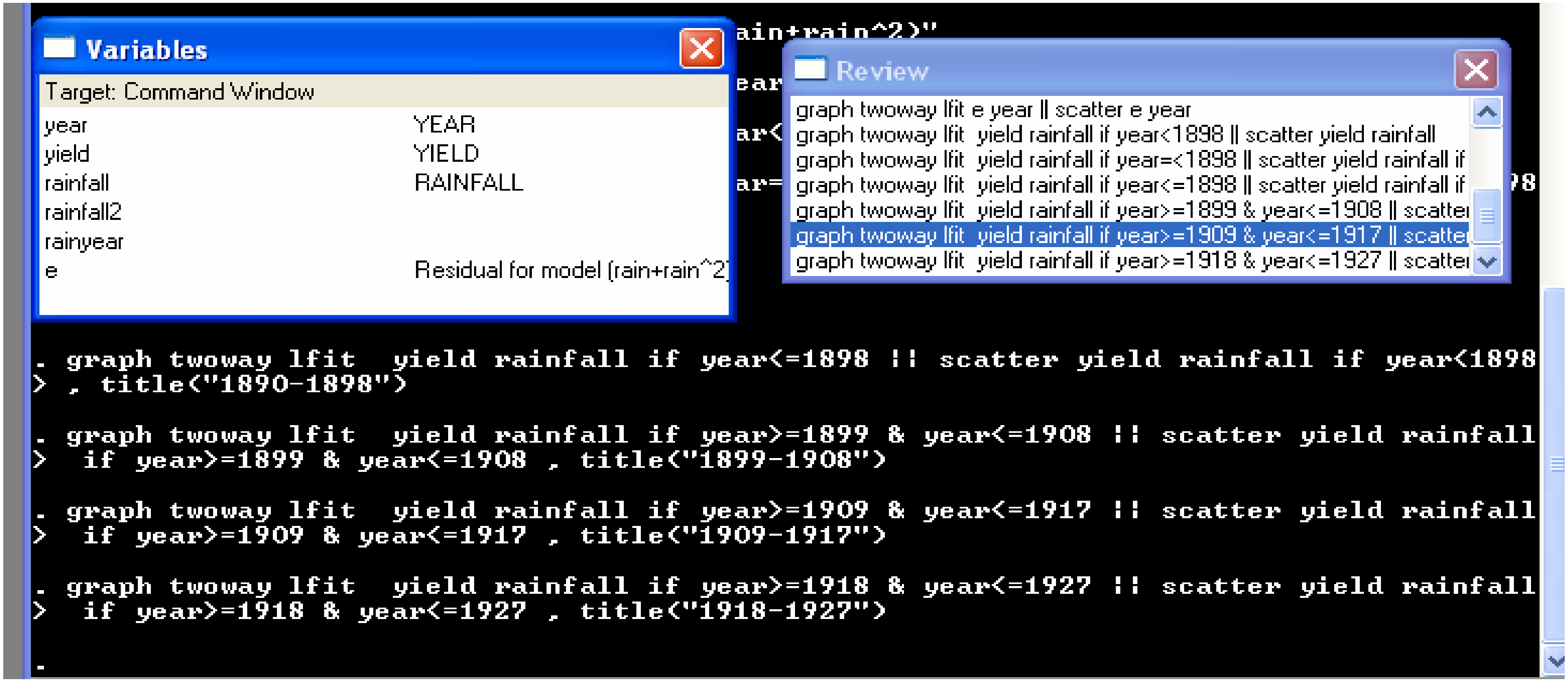


Partly because of the outliers and
partly because we suspect that the
effect of rain might be changing over
1890 to 1928 (because of improvements in
agricultural techniques, including irrigation),

it seems appropriate to further investigate
the interactive effect
of year and rainfall on yield.

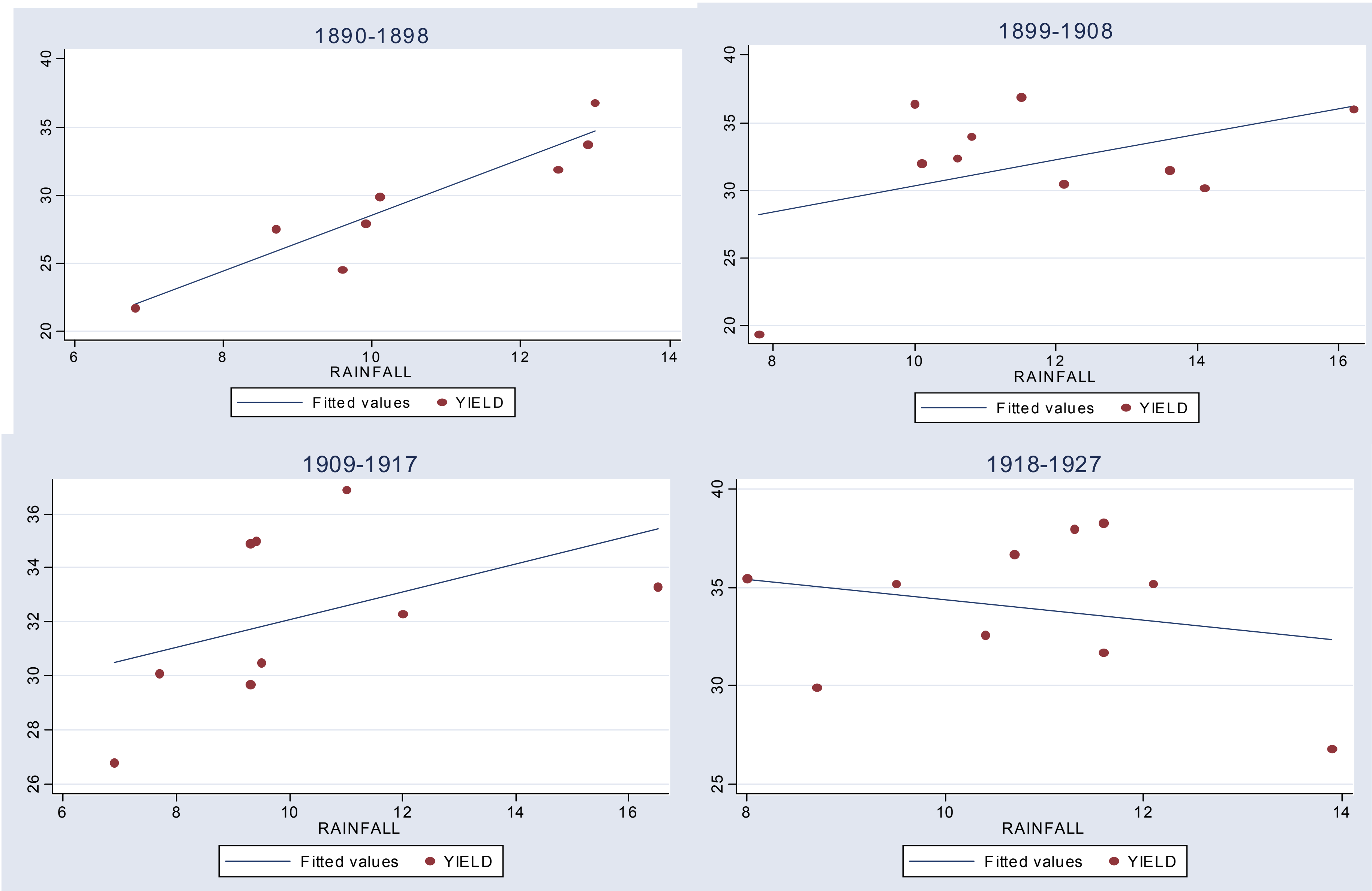
Conditional scatter plots:

STATA commands



Note: The conditional scatterplots show the effect of rainfall on yield to be smaller in later time periods .

Conditional scatter plots



Fitted Model

Final regression model with quadratic functions and interaction terms

$$Yield=\beta_0+ \beta_1 rainfall+ \beta_2 rainfall^2+ \beta_3 Year+ \beta_3(Rainfall*Year)$$

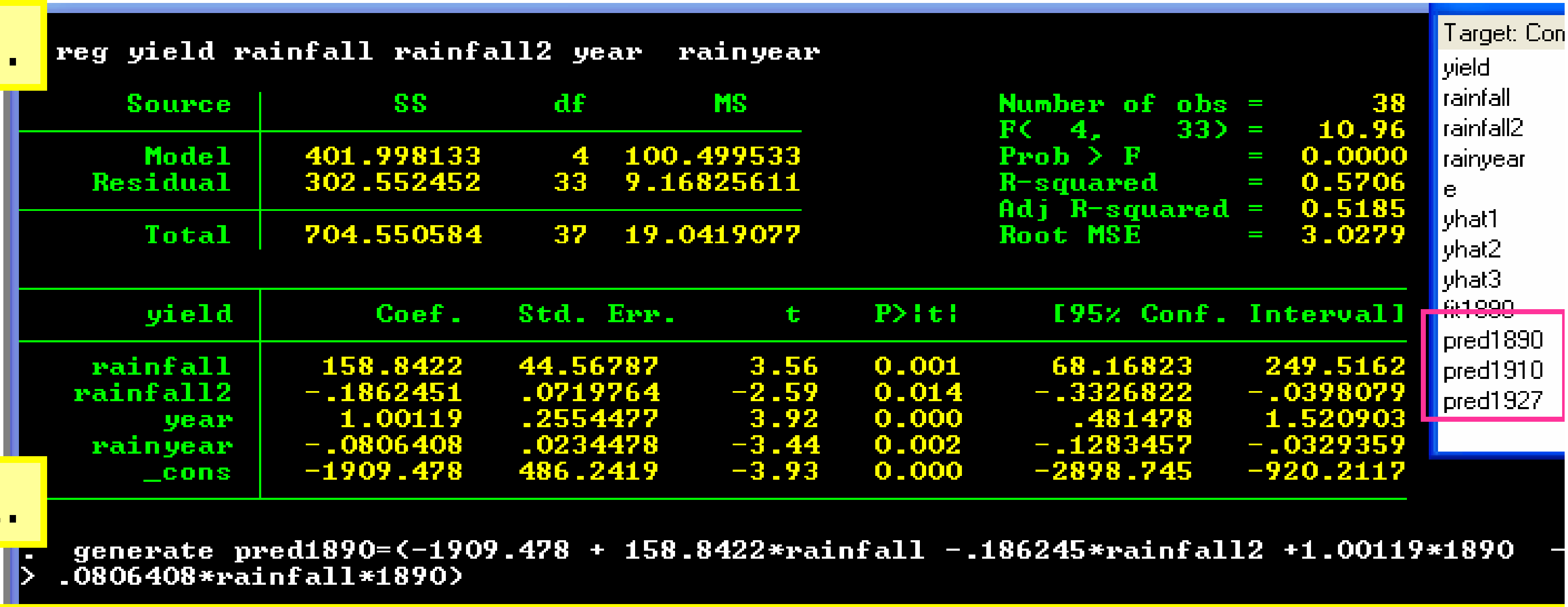
```
reg yield rainfall rainfall2 year  rainyear

      Source |       SS       df       MS              Number of obs =      38
-------------+------------------------------           F(  4,    33) =   10.96
       Model |  401.998133     4  100.499533           Prob > F      =  0.0000
    Residual |  302.552452    33  9.16825611           R-squared     =  0.5706
-------------+------------------------------           Adj R-squared =  0.5185
       Total |  704.550584    37  19.0419077           Root MSE      =  3.0279

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rainfall |   158.8422   44.56787     3.56   0.001     68.16823    249.5162
   rainfall2 |  -.1862451   .0719764    -2.59   0.014    -.3326822   -.0398079
        year |    1.00119   .2554477     3.92   0.000      .481478    1.520903
    rainyear |  -.0806408   .0234478    -3.44   0.002    -.1283457   -.0329359
       _cons |  -1909.478   486.2419    -3.93   0.000    -2898.745   -920.2117
------------------------------------------------------------------------------
```

Quadratic regression lines for 1890, 1910 & 1927

$$Yield=\beta_0+\beta_1 rainfall+\beta_2 rainfall^2+\beta_3 Year+\beta_3 (Rainfall*Year)$$

1. Run the regression
2. Use the regression estimates and substitute the corresponding year in the model to generate 3 new variables:
The predicted yields for year=1890,1910,1927

1.



```
reg yield rainfall rainfall2 year  rainyear

      Source |       SS       df       MS              Number of obs =      38
-------------+------------------------------           F(  4,    33) =   10.96
       Model |  401.998133     4  100.499533           Prob > F      =  0.0000
    Residual |  302.552452    33  9.16825611           R-squared     =  0.5706
-------------+------------------------------           Adj R-squared =  0.5185
       Total |  704.550584    37  19.0419077           Root MSE      =  3.0279

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rainfall |   158.8422   44.56787     3.56   0.001     68.16823    249.5162
   rainfall2 |  -.1862451   .0719764    -2.59   0.014    -.3326822   -.0398079
        year |    1.00119   .2554477     3.92   0.000      .481478    1.520903
    rainyear |  -.0806408   .0234478    -3.44   0.002    -.1283457   -.0329359
       _cons |  -1909.478   486.2419    -3.93   0.000    -2898.745   -920.2117
------------------------------------------------------------------------------

. generate pred1890=(-1909.478 + 158.8422*rainfall -.186245*rainfall2 +1.00119*1890  -
> .0806408*rainfall*1890)
```


2.

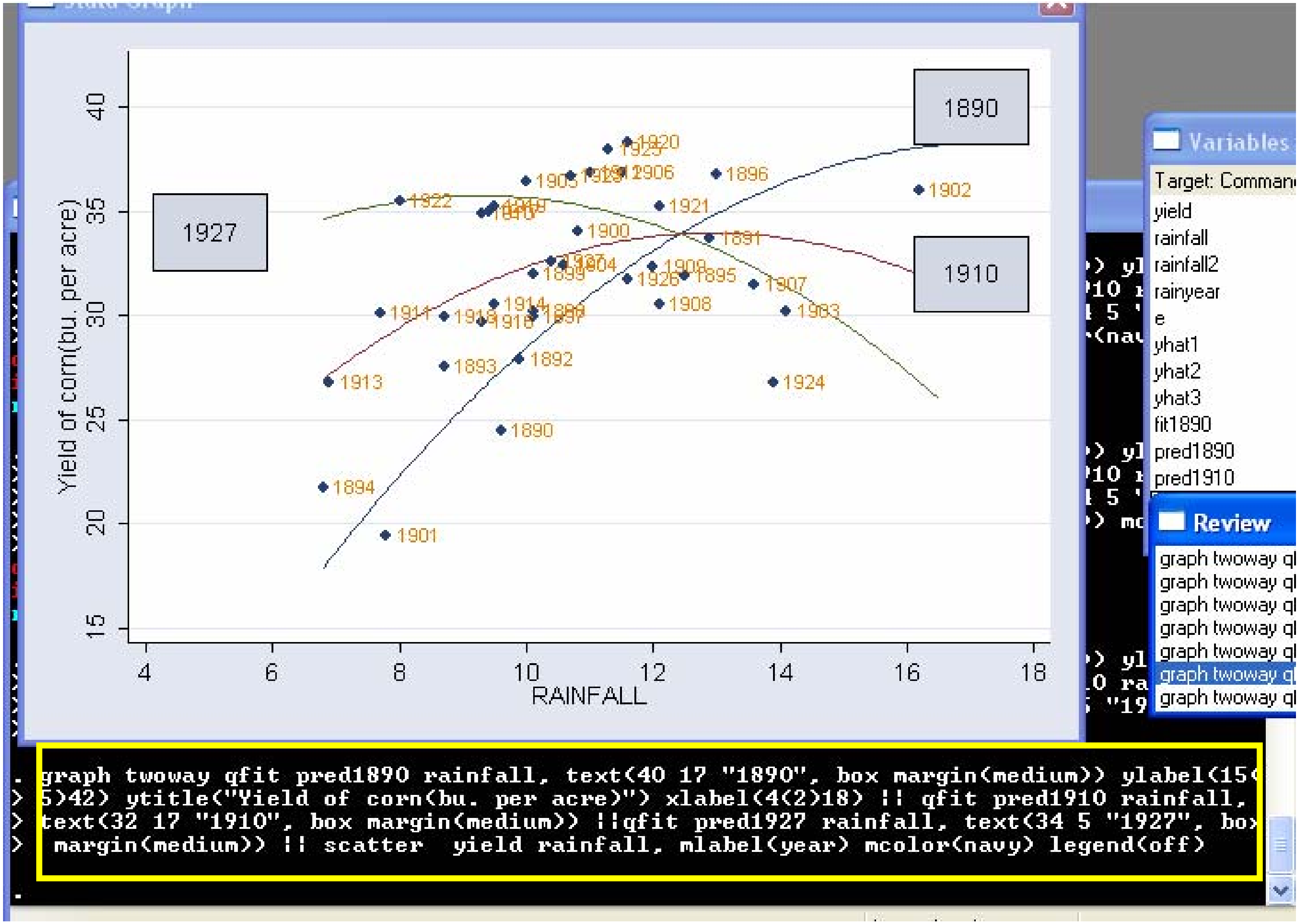
$$Pred1890=\beta_0+\beta_1 rainfall+\beta_2 rainfall^2+\beta_3 1890+\beta_3 (Rainfall*1890)$$

The predicted yield values generated for years: 1890, 1910 and 1927

```
Stata Results

. reg yield rainfall rainfall2 year  rainyear

      Source |       SS       df       MS              Number of obs =      38
-------------+------------------------------           F(  4,    33) =   10.96
       Model |  401.998133     4  100.499533           Prob > F      =  0.0000
    Residual |  302.552452    33  9.16825611           R-squared     =  0.5706
-------------+------------------------------           Adj R-squared =  0.5185
       Total |  704.550584    37  19.0419077           Root MSE      =  3.0279

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rainfall |   158.8422   44.56787     3.56   0.001     68.16823    249.5162
   rainfall2 |  -.1862451   .0719764    -2.59   0.014    -.3326822   -.0398079
        year |    1.00119   .2554477     3.92   0.000      .481478    1.520903
    rainyear |  -.0806408   .0234478    -3.44   0.002    -.1283457   -.0329359
       _cons |  -1909.478   486.2419    -3.93   0.000    -2898.745   -920.2117
------------------------------------------------------------------------------

.  generate pred1890=(-1909.478 + 158.8422*rainfall -.186245*rainfall2 +1.00119*1890  -
> .0806408*rainfall*1890)

.  generate pred1910=(-1909.478 + 158.8422*rainfall -.186245*rainfall2 +1.00119*1910  -
> .0806408*rainfall*1910)

.  generate pred1927=(-1909.478 + 158.8422*rainfall -.186245*rainfall2 +1.00119*1927  -
> .0806408*rainfall*1927)
```

```
Varial
Target: Con
yield
rainfall
rainfall2
rainyear
e
yhat1
yhat2
yhat3
fit1890
pred1890
pred1910
pred1927
```

Yearly corn yield vs rainfall between 1890 and 1927 and quadratic regression lines for years 1890, 1910 and 1927



Summary of Findings

•As evident in the scatterplot above, the mean yearly yield of corn in six Midwestern states from 1890 to 1927 increased with increasing rainfall up to a certain optimum rainfall, and then leveled off or decreased with rain in excess of that amount (the p-value from a t-test for the quadratic effect of rainfall on mean corn yield is .014).

•There is strong evidence, however, that the effect of rainfall changed over this period of observation (p-value from a t-test for the interactive effect of year and rainfall is .002).

•Representative quadratic fits to the regression of corn yield on rainfall are shown in the plot—for 1890, 1910, and 1927. It is apparent that less rainfall was needed to produce the same mean yield as time progressed.

Example:
Causes of Student Academic Performance

- Randomly sampling 400 elementary schools from the California Department of Education's API 2000 dataset.
- Data contains a measure of school academic performance as well as other attributes of the elementary schools, such as, class size, enrollment, poverty, etc.
- See Handout…