Lecture 2
Linear Regression: A Model for the Mean

Sharyn O'Halloran

Closer Look at:

- Linear Regression Model
  - Least squares procedure
  - Inferential tools
  - Confidence and Prediction Intervals
- Assumptions
- Robustness
- Model checking
- Log transformation (of *Y*, *X*, or both)

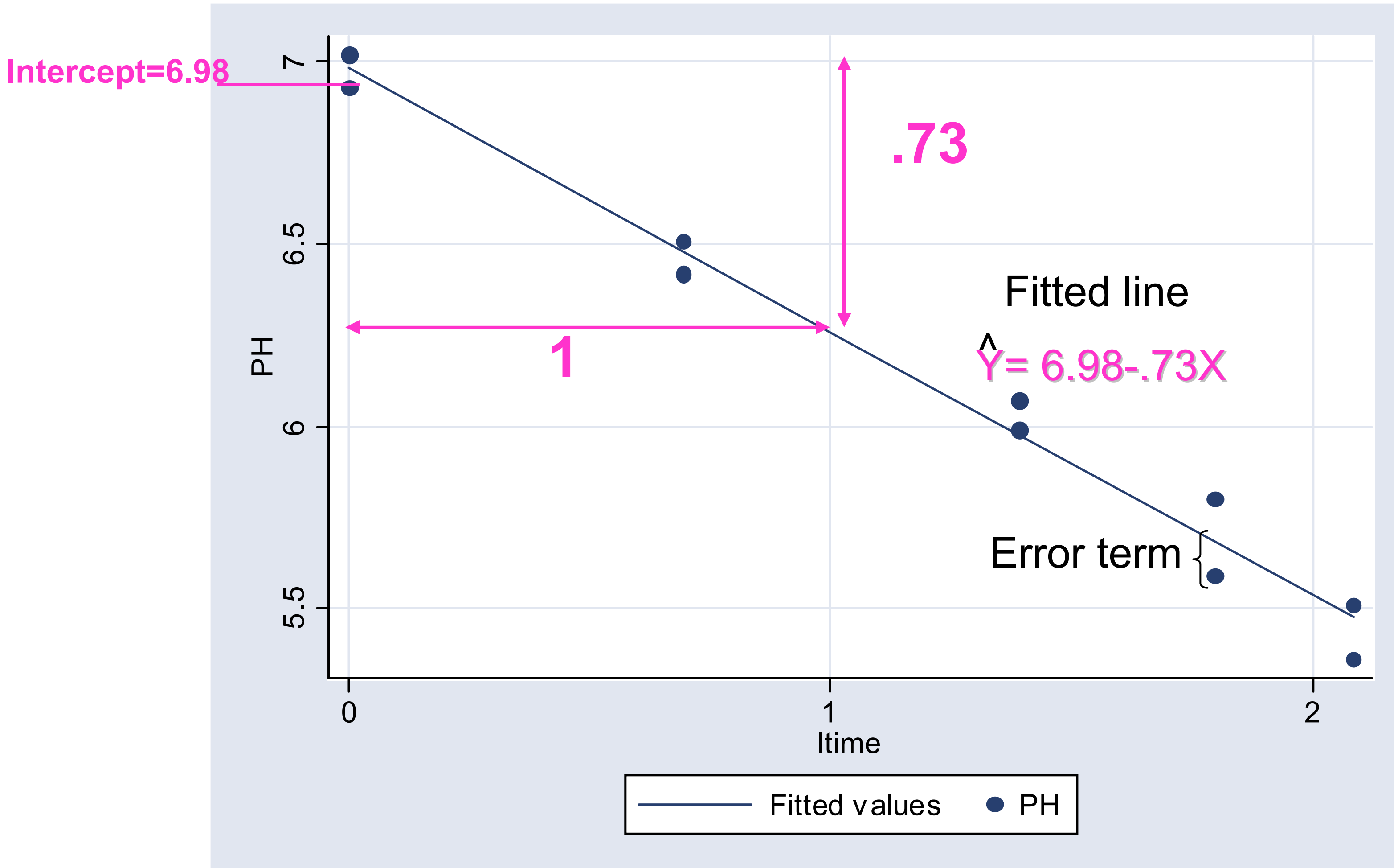
Linear Regression: Introduction

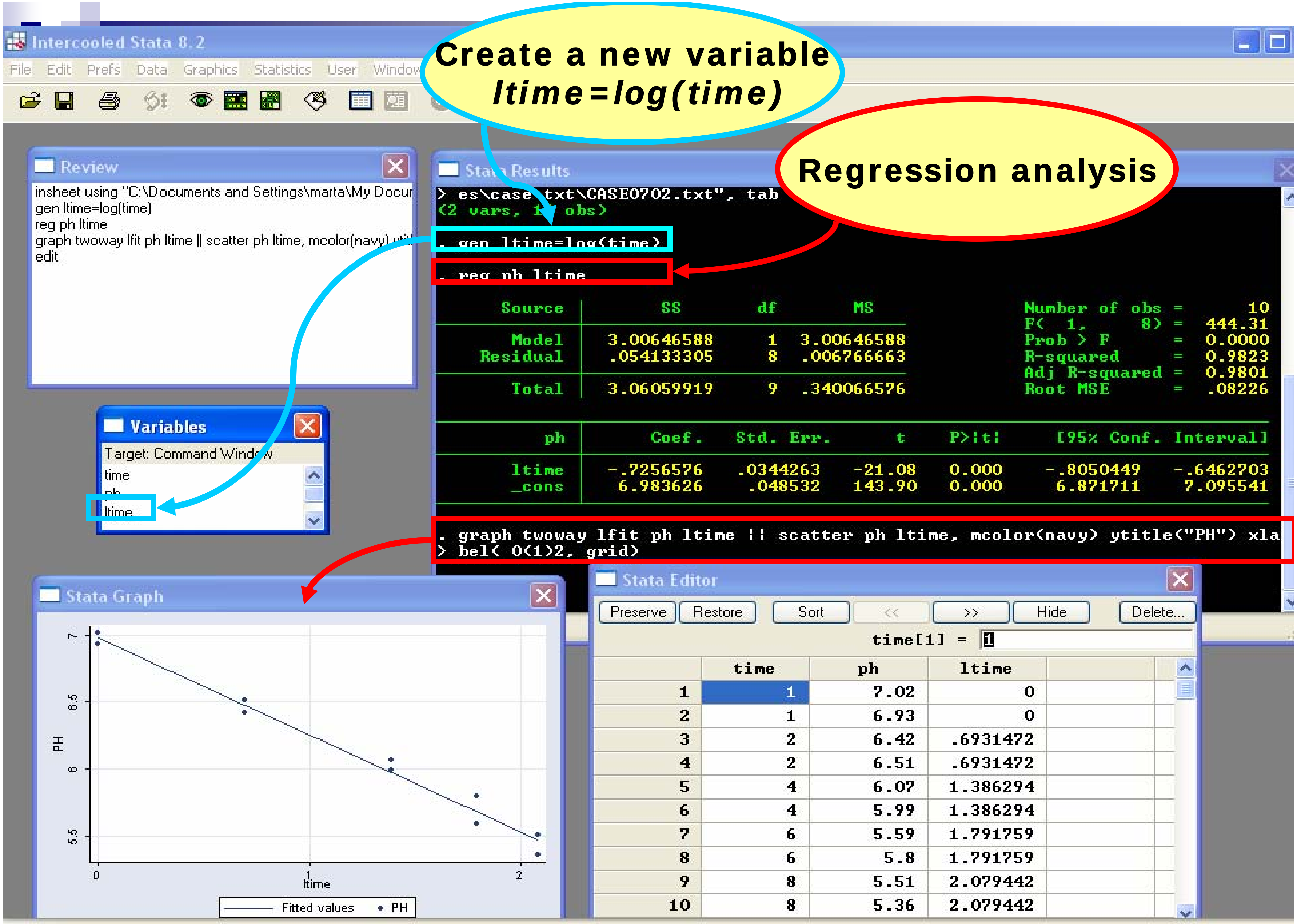
- Data: (Y_i, X_i) for i = 1,…,n
- Interest is in the probability distribution of *Y* as a function of *X*
- Linear Regression model:
  - Mean of *Y* is a straight line function of *X*, *plus an error term or residual*
  - *Goal is to find the best fit line that minimizes the sum of the error terms*

Estimated regression line

Steer example (see Display 7.3, p. 177)

Equation for estimated regression line:




Create a new variable
ltime=log(time)
Regression analysis
Intercooled Stata 8.2
File Edit Prefs Data Graphics Statistics User Window
Review
insheet using "C:\Documents and Settings\marta\My Docur
gen ltime=log(time)
reg ph ltime
graph twoway lfit ph ltime || scatter ph ltime, mcolor(navy) yti
edit
Variables
Target: Command Window
time
ltime
Stata Results
> es\case txt\CASE0702.txt", tab
(2 vars, 10 obs)
. gen ltime=log(time)
. reg ph ltime
Source SS df MS
Model 3.00646588 1 3.00646588
Residual .054133305 8 .006766663
Total 3.06059919 9 .340066576
Number of obs = 10
F(1, 8) = 444.31
Prob > F = 0.0000
R-squared = 0.9823
Adj R-squared = 0.9801
Root MSE = .08226
ph Coef. Std. Err. t P>|t| [95% Conf. Interval]
ltime -.7256576 .0344263 -21.08 0.000 -.8050449 -.6462703
_cons 6.983626 .048532 143.90 0.000 6.871711 7.095541
. graph twoway lfit ph ltime || scatter ph ltime, mcolor(navy) ytitle("PH") xla
> bel(0(1)2, grid)
Stata Graph
PH
7
6.5
6
5.5
0
1
2
ltime
Fitted values
PH
Stata Editor
Preserve Restore Sort << >> Hide Delete...
time[1] = 1
time ph ltime
1 1 7.02 0
2 1 6.93 0
3 2 6.42 .6931472
4 2 6.51 .6931472
5 4 6.07 1.386294
6 4 5.99 1.386294
7 6 5.59 1.791759
8 6 5.8 1.791759
9 8 5.51 2.079442
10 8 5.36 2.079442

Regression Terminology

Regression: the mean of a response variable as a function of one or more explanatory variables:

$$\mu\{Y \mid X\}$$

Regression model: an ideal formula to approximate the regression

Simple linear regression model:

$$\mu\{Y \mid X\} = \beta_0 + \beta_1 X$$

"mean of Y given X" or "regression of Y on X"

Intercept

Slope

Unknown parameter

Regression Terminology

Y	X
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable

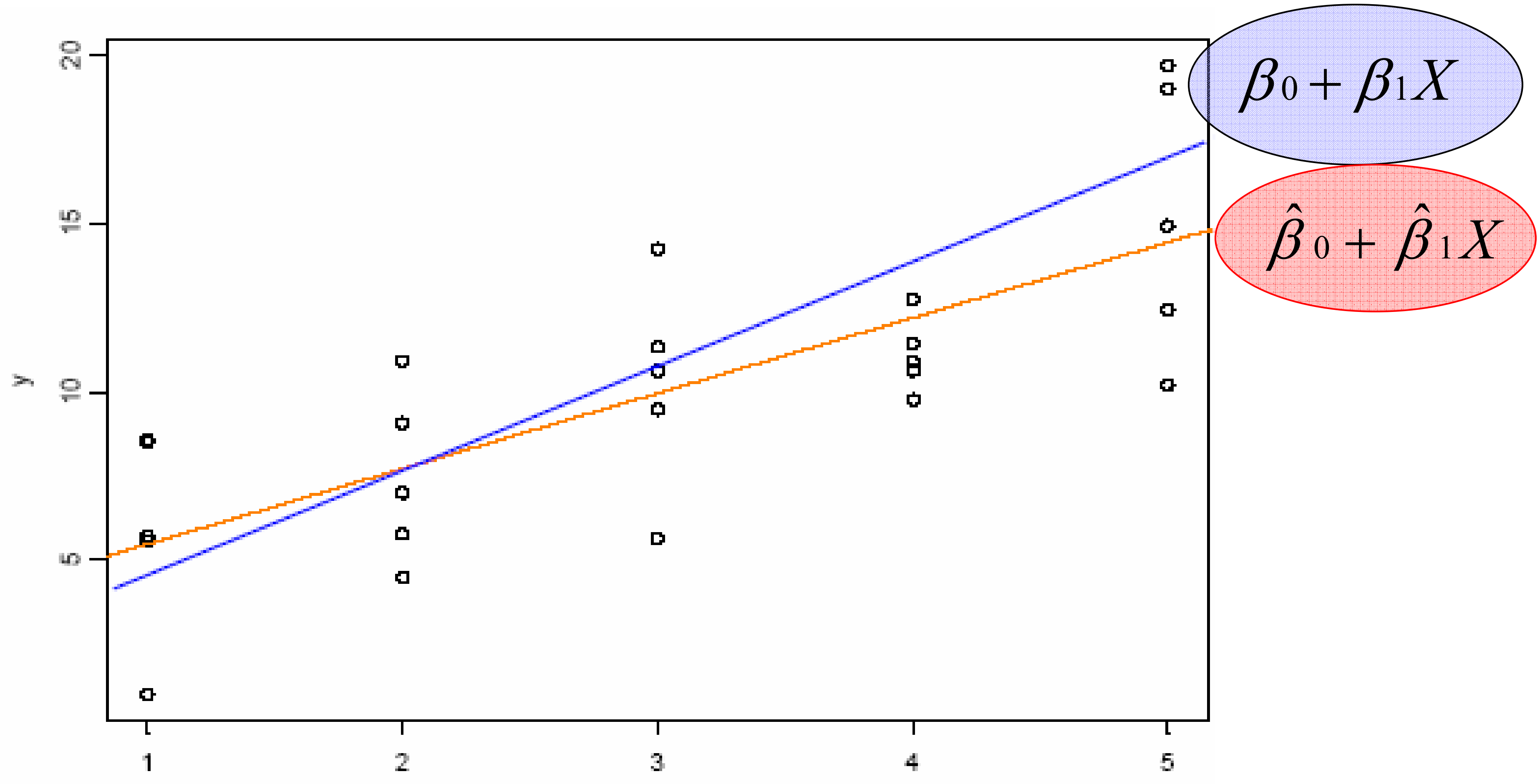
Y's probability distribution is to be explained by X

b_0 and b_1 are the regression coefficients

(See Display 7.5, p. 180)

Note: $Y = b_0 + b_1 X$ is NOT simple regression

Regression Terminology: Estimated coefficients



Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the residuals small

Regression Terminology

- Fitted value for obs. i is its estimated mean: $\hat{Y} = fit_i = \mu\{Y \mid X\} = \beta_0 + \beta_1 X$
- Residual for obs. i:

$$\mathrm{res}_{\mathrm{i}} = \mathrm{Y}_{\mathrm{i}} - \mathrm{fit}_{\mathrm{i}} \Rightarrow e_i = Y_i - \hat{Y}$$

- Least Squares statistical estimation method finds those estimates that minimize the sum of squared residuals.

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Solution (from calculus) on p. 182 of Sleuth

Least Squares Procedure

- **The Least-squares procedure obtains estimates of the linear equation coefficients β_0 and β_1, in the model**

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- **by minimizing the sum of the squared residuals or errors (e_i)**

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- **This results in a procedure stated as**

$$SSE = \sum e_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Choose β_0 and β_1 so that the quantity is minimized.

Least Squares Procedure

- **The slope coefficient estimator is**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} = r_{xy}\frac{s_Y}{s_X}$$

CORRELATION BETWEEN X AND Y

STANDARD DEVIATION OF Y OVER THE STANDARD DEVIATION OF X

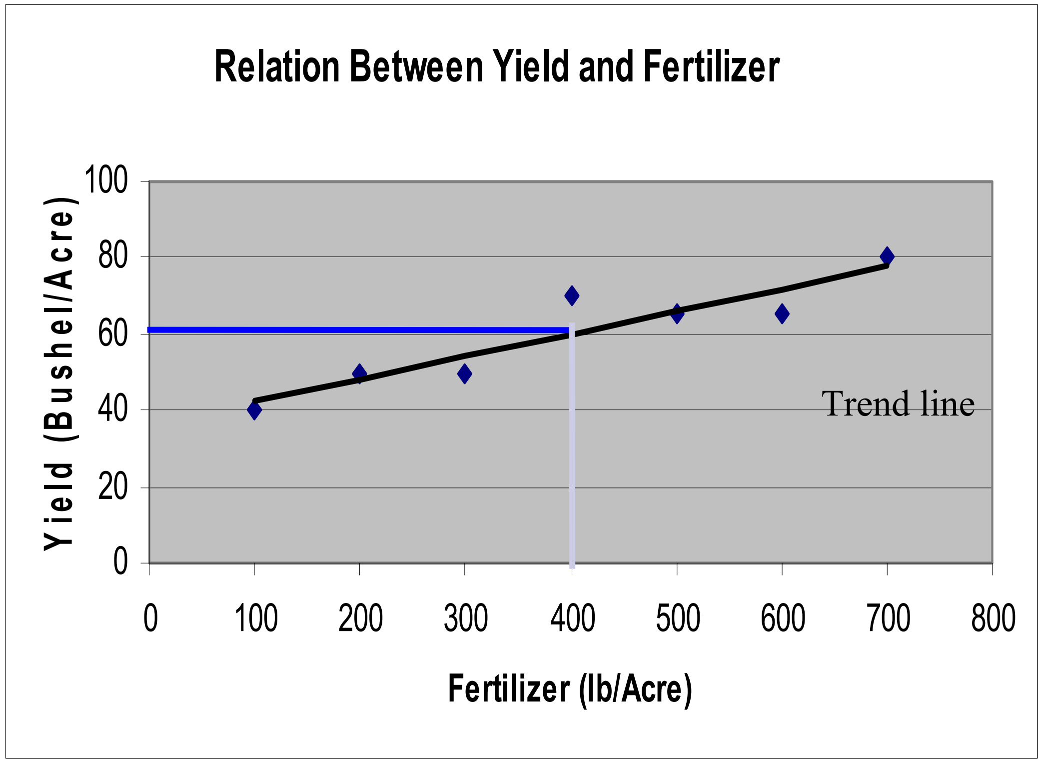
- **And the constant or intercept indicator is**

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

Least Squares Procedure(cont.)

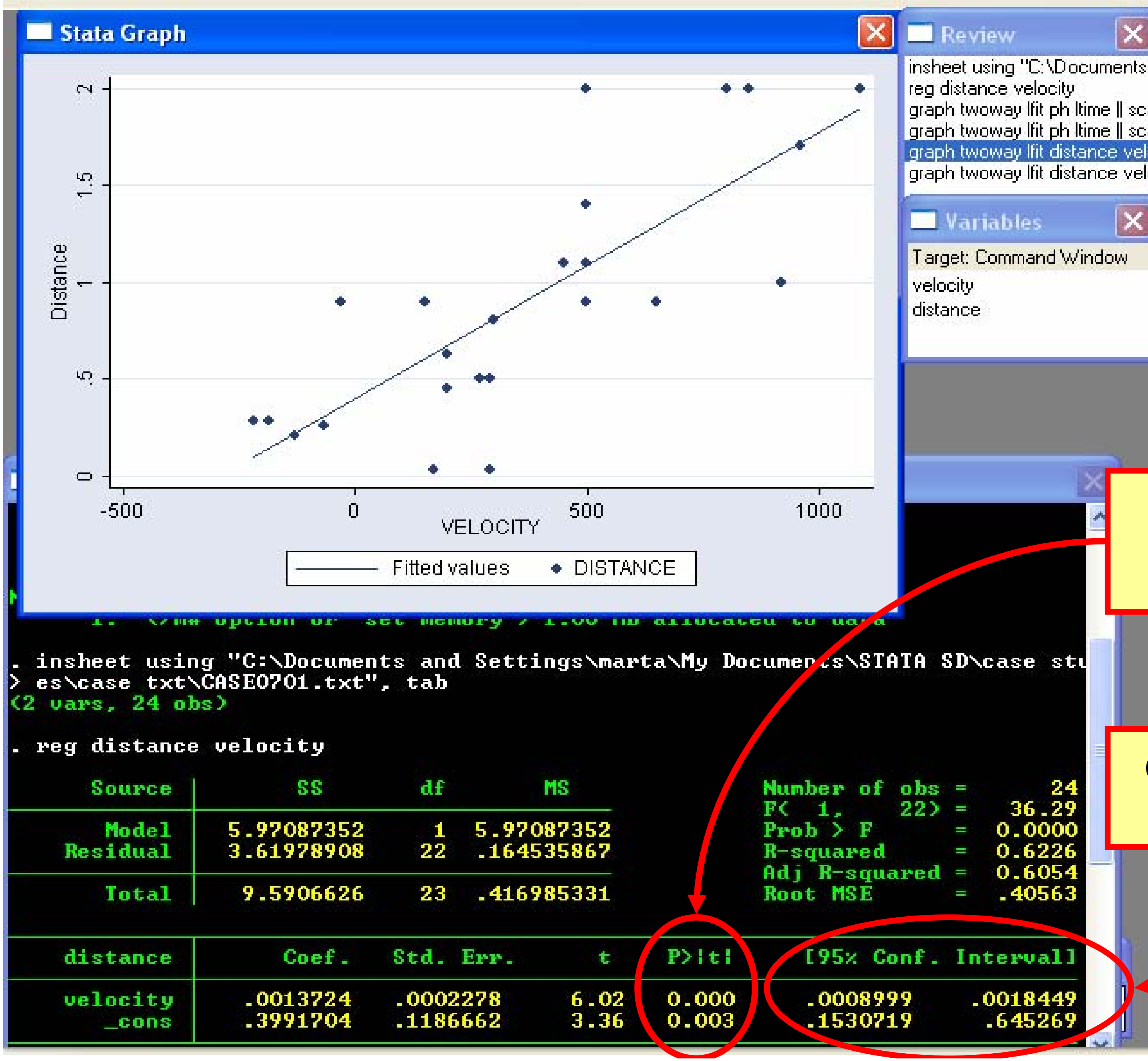
- Note that the regression line always goes through the mean X, Y.
- Think of this regression line as the expected value of Y for a given value of X.

That is, for any value of the independent variable there is a single most likely value for the dependent variable



Tests and Confidence Intervals for β_0, β_1

- Degrees of freedom:
  - (n-2) = sample size - number of coefficients
- Variance {Y|X}
  - σ^2= (sum of squared residuals)/(n-2)
- Standard errors (p. 184)
- Ideal normal model:
  - the sampling distributions of β_0 and β_1 have the shape of a t-distribution on (n-2) d.f.
- Do t-tests and CIs as usual (df=n-2)



P values for $H_o=0$

Confidence intervals

Inference Tools

- Hypothesis Test and Confidence Interval for mean of Y at some X:
  - Estimate the mean of *Y* at $X = X_0$ by

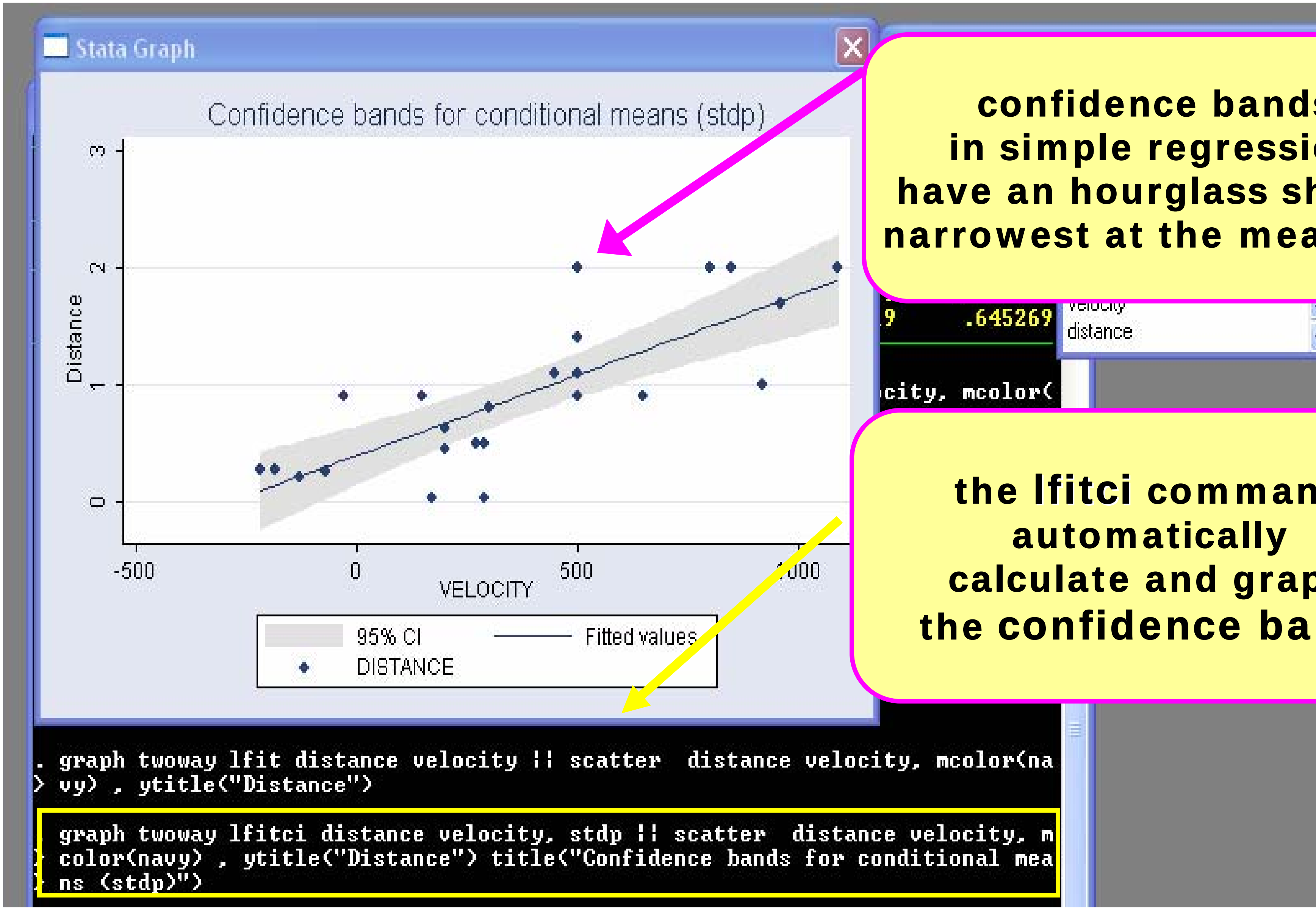
$$\hat{\mu}\{Y \mid X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

  - Standard Error of $\hat{\beta}_0$

$$SE[\hat{\mu}\{Y \mid X_0\}] = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_x^2}}$$

- Conduct t-test and confidence interval in the usual way (df = n-2)

Confidence bands for conditional means



confidence bands in simple regression have an hourglass shape, narrowest at the mean of X

the lfitci command automatically calculate and graph the confidence bands

Prediction

- Prediction of a future Y at $X=X_0$

$$\text{Pred}(Y \mid X_0) = \hat{\mu}\{Y \mid X_0\}$$

- Standard error of prediction:

$$SE[\text{Pred}(Y \mid X_0)] = \sqrt{\hat{\sigma}^2 + (SE[\hat{\mu}(Y \mid X_0)])^2}$$

Variability of Y about its mean

Uncertainty in the estimated mean

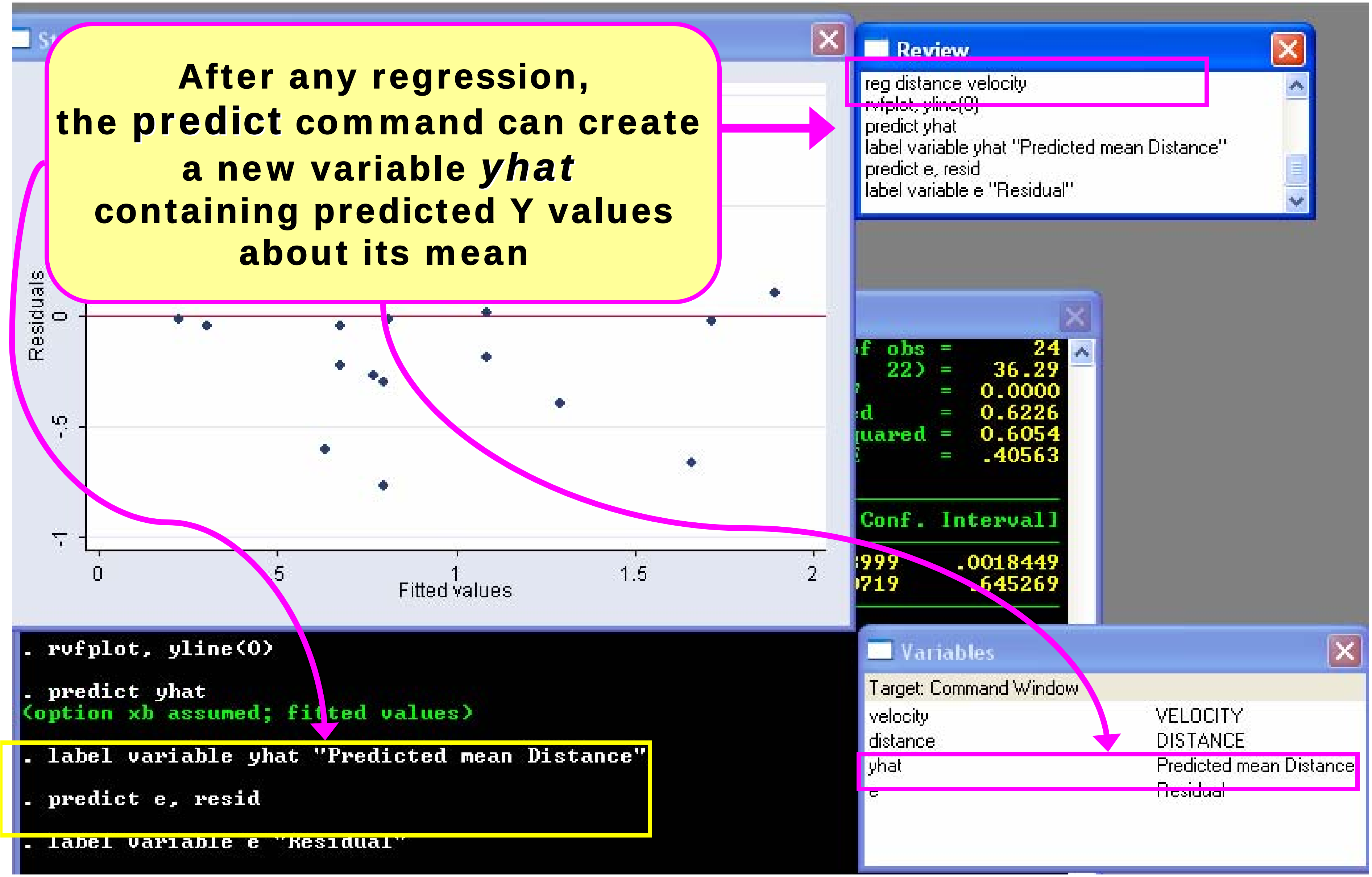
- 95% prediction interval:

$$\text{Pred}(Y \mid X_0) \pm t_{df}(.975) * SE[\text{Pred}(Y \mid X_0)]$$

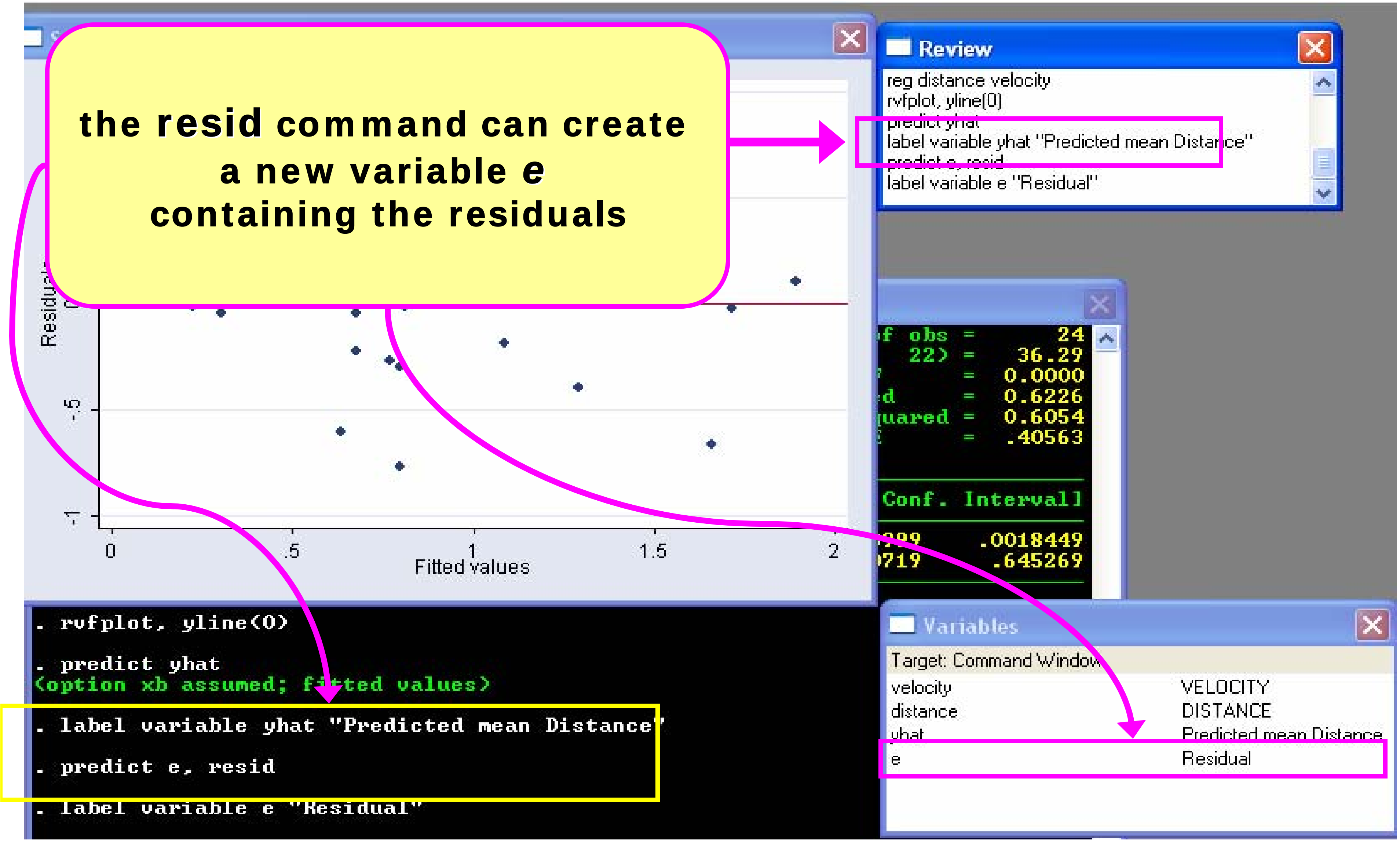
Residuals vs. predicted values plot

After any regression analysis we can automatically draw a residual-versus-fitted plot just by typing

Predicted values (*yhat*)

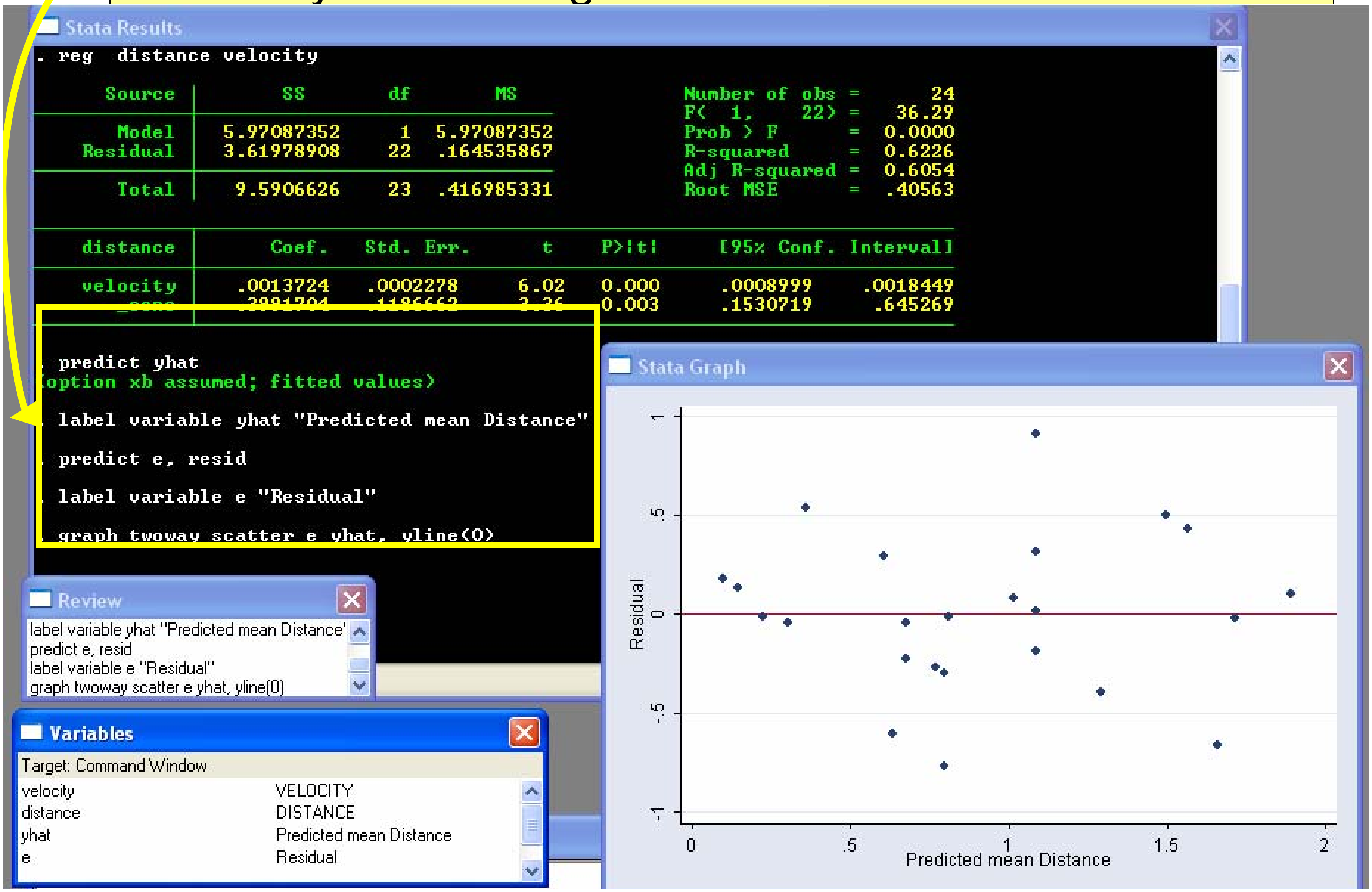


Residuals (*e*)

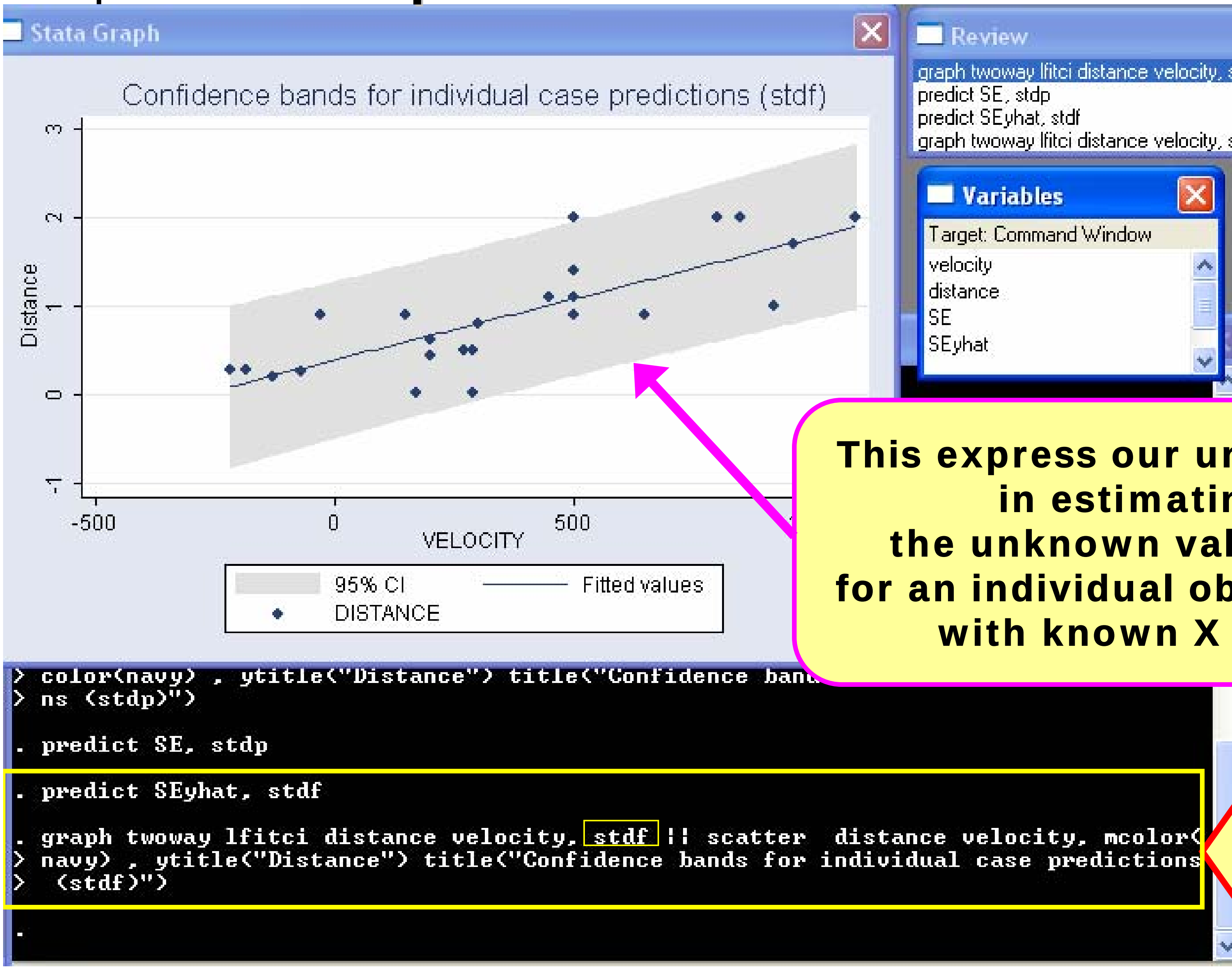


the resid command can create a new variable e containing the residuals
Review
reg distance velocity
rvfplot, yline(0)
predict yhat
label variable yhat "Predicted mean Distance"
predict e, resid
label variable e "Residual"
Residuals
Fitted values
Conf. Interval]
.0018449
.645269
Variables
Target: Command Window
velocity VELOCITY
distance DISTANCE
yhat Predicted mean Distance
e Residual
. rvfplot, yline(0)
. predict yhat
(option xb assumed; fitted values)
. label variable yhat "Predicted mean Distance"
. predict e, resid
. label variable e "Residual"

The residual-versus-predicted-values plot could be drawn “by hand” using these commands



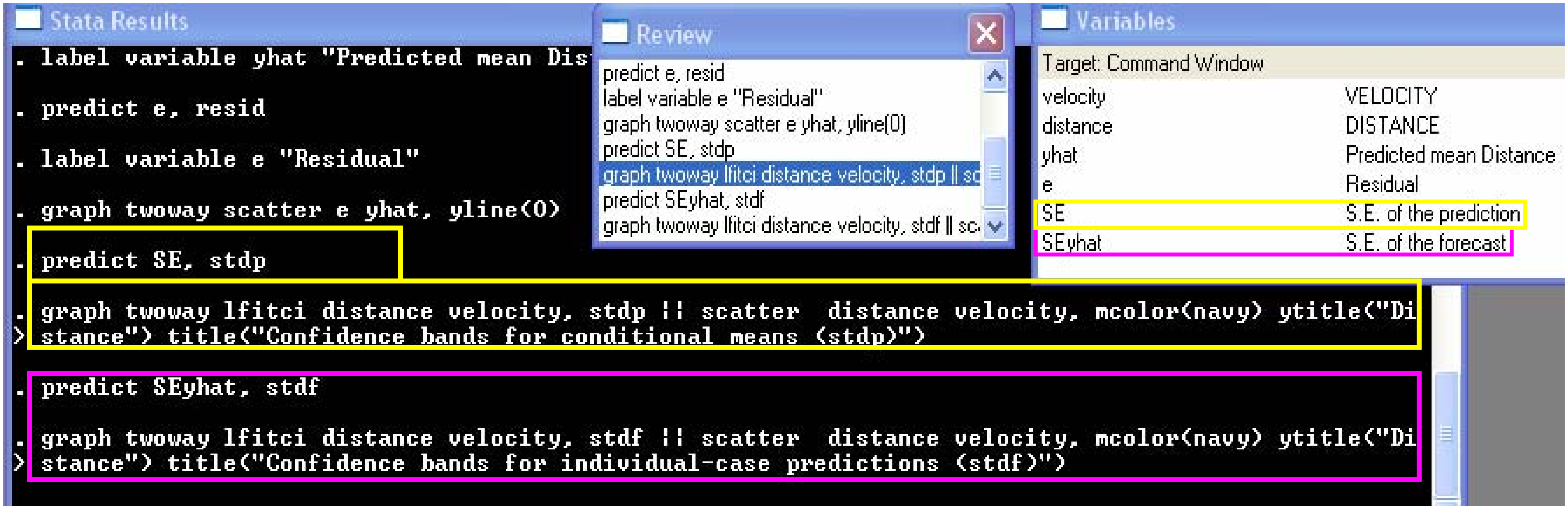
Second type of confidence interval for regression prediction: **“prediction band”**

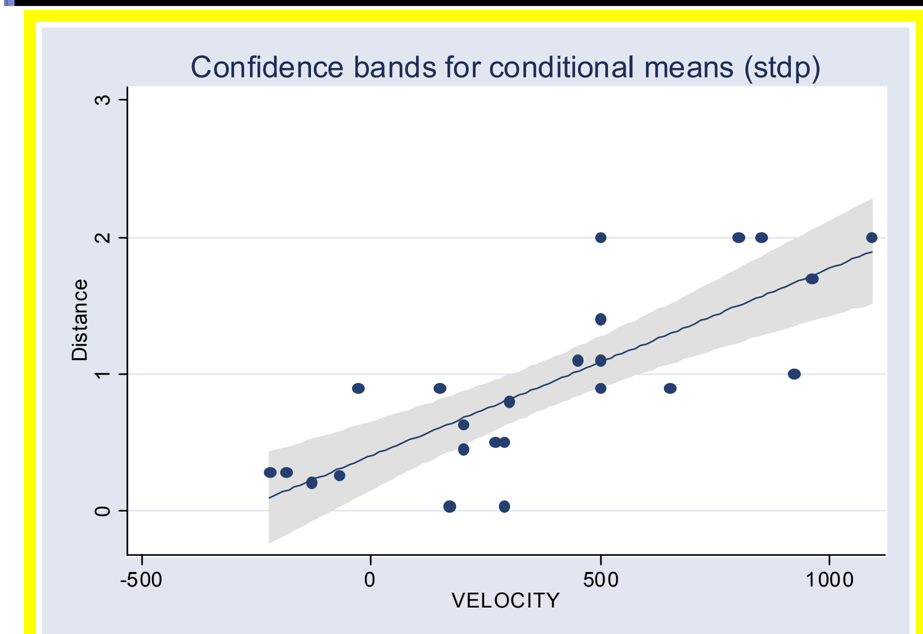


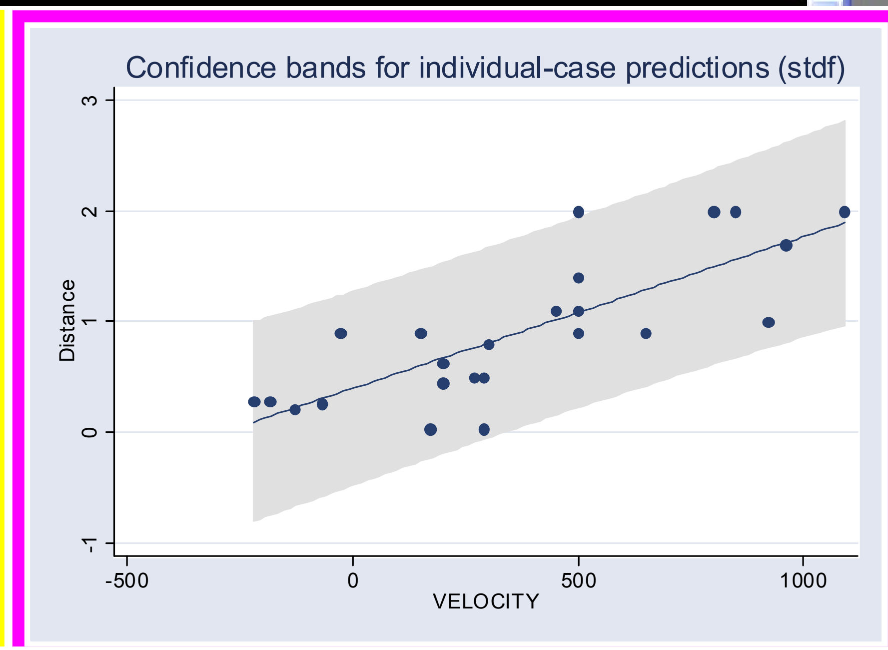
This express our uncertainty in estimating the unknown value of Y for an individual observation with known X value

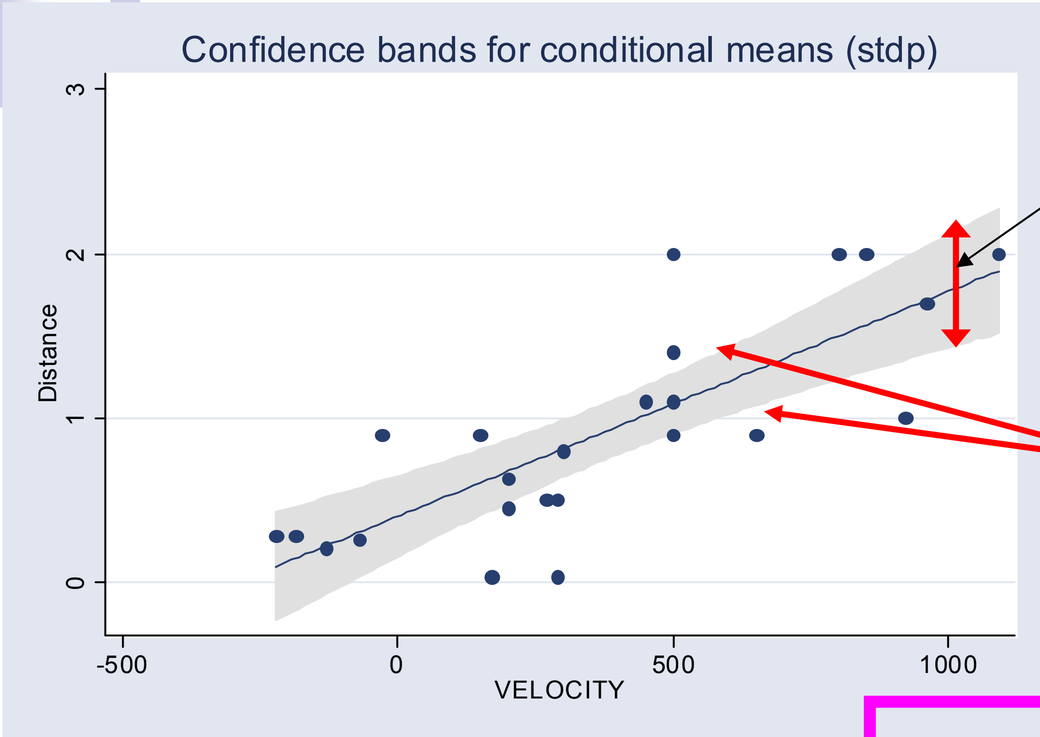
Command: **lftci** with **stdf** option

Additional note: Predict can generate two kinds of standard errors for the predicted y value, which have two different applications.



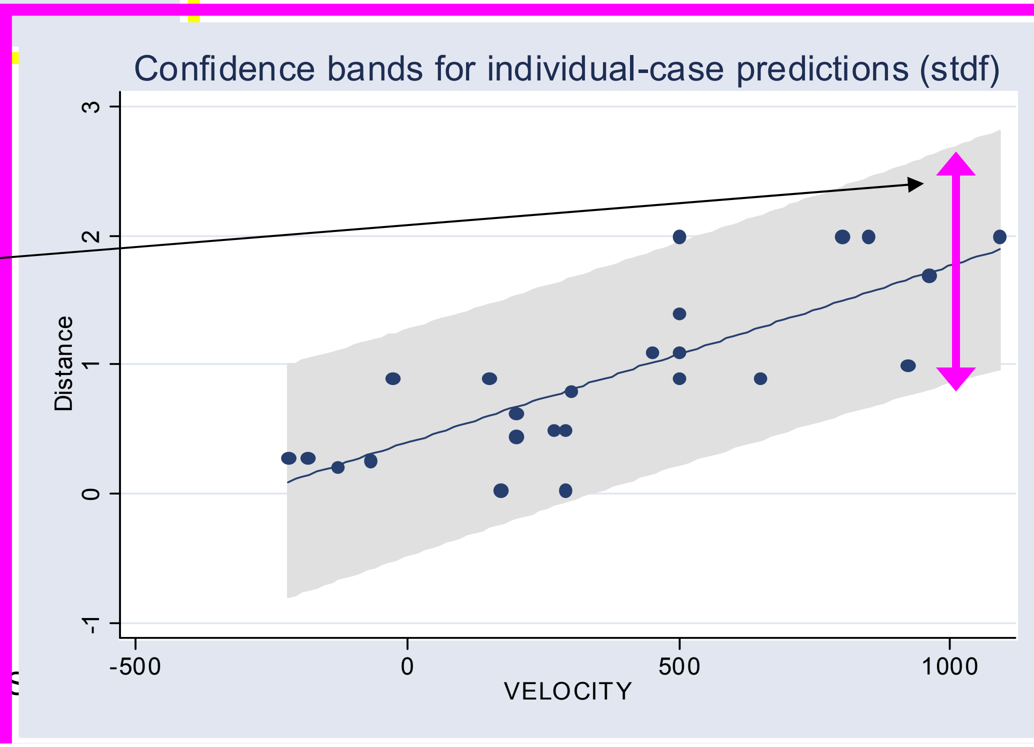






95% confidence interval
for $\mu\{Y|1000\}$

confidence band:
a set of
confidence intervals
for $\mu\{Y|X_0\}$

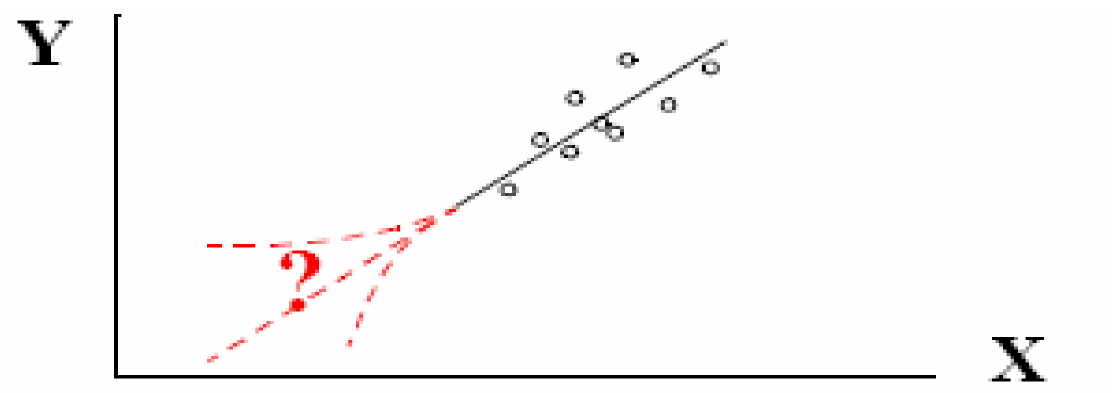


95% prediction interval
for Y at $X=1000$

Calibration interval:
values of $\mathbf{X}$ for which $\mathbf{Y_0}$ is in a prediction interval

Notes about confidence and prediction bands

- Both are narrowest at the mean of *X*
- Beware of *extrapolation*



- The width of the Confidence Interval is zero if n is large enough; this is not true of the Prediction Interval.

Review of simple linear regression

1. Model with **constant variance.**

$$\mu\{Y \mid X\} = \beta_0 + \beta_1 X$$

$$\text{var}\{Y \mid X\} = \sigma^2$$

2. **Least squares**: choose estimators β_0 and β_1 to minimize the sum of squared residuals.

$$\hat{\beta}_1 = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) / \sum_{i=1}^{n}(X_i - \overline{X})^2.$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$res_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \, (i = 1,..,n)$$

3. **Properties** of estimators.

$$\hat{\sigma}^2 = \sum_{i=1}^{n} res_i^{\,2} / (n-2)$$

$$SE(\hat{\beta}_1) = \hat{\sigma} / \sqrt{(n-1)s_x^2}$$

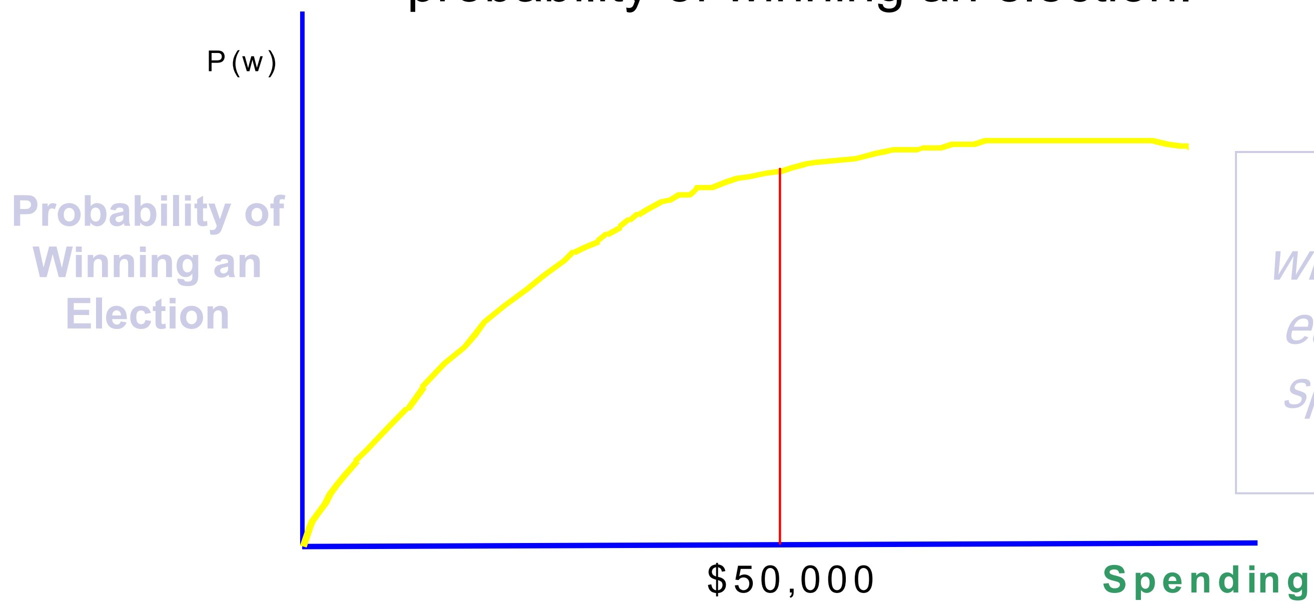
$$SE(\hat{\beta}_0) = \hat{\sigma} / \sqrt{(1/n) + \overline{X}^2 / (n-1)s_x^2}$$

Assumptions of Linear Regression

- A linear regression model assumes:
  - Linearity:
    - $\mu\{Y|X\} = \beta_0 + \beta_1 X$
  - Constant Variance:
    - $\text{var}\{Y|X\} = \sigma^2$
  - Normality
    - Dist. of Y's at any X is normal
  - Independence
    - Given X_i's, the Y_i's are independent

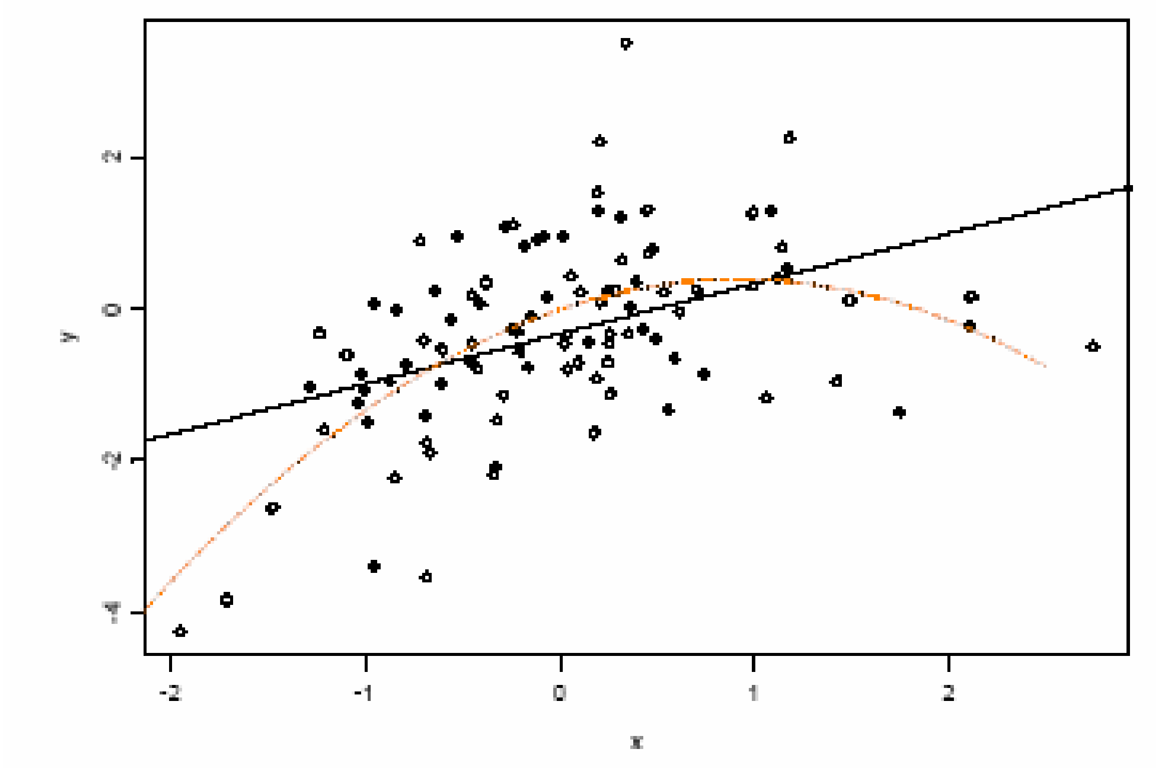
Examples of Violations

- Non-Linearity
  - The true relation between the independent and dependent variables may not be linear.
    - For example, consider campaign fundraising and the probability of winning an election.



Consequences of violation of linearity

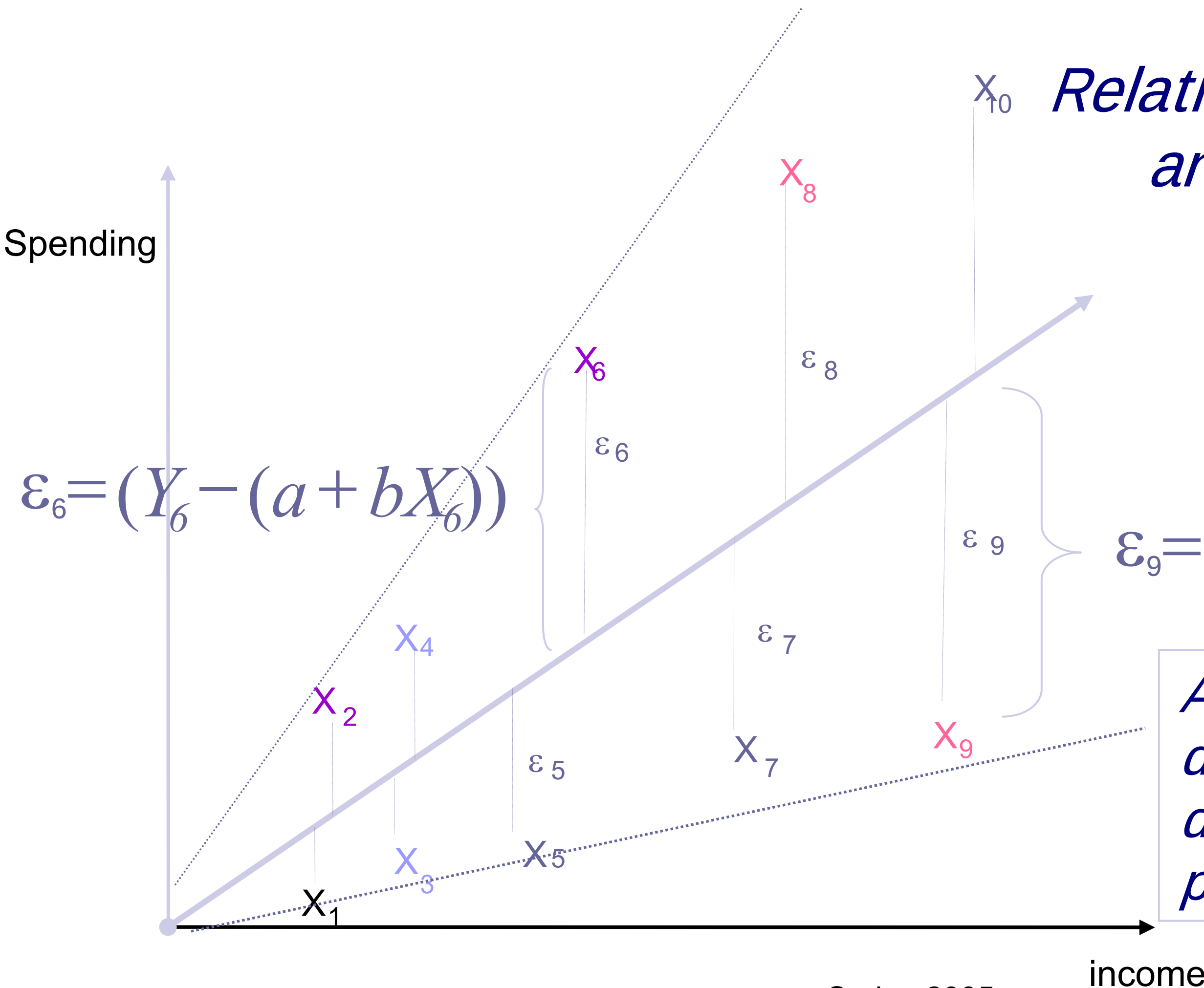
- If "linearity" is violated, misleading conclusions may occur (however, the degree of the problem depends on the degree of non-linearity)



Examples of Violations: Constant Variance

- Constant Variance or Homoskedasticity
  - The Homoskedasticity assumption implies that, on average, we do *not expect* to get larger errors in some cases than in others.
    - Of course, due to the luck of the draw, some errors will turn out to be larger then others.
    - But homoskedasticity is violated only when this happens in a predictable manner.
  - Example: income and spending on certain goods.
    - People with higher incomes have more choices about what to buy.
    - We would expect that there consumption of certain goods is more variable than for families with lower incomes.

Violation of constant variance



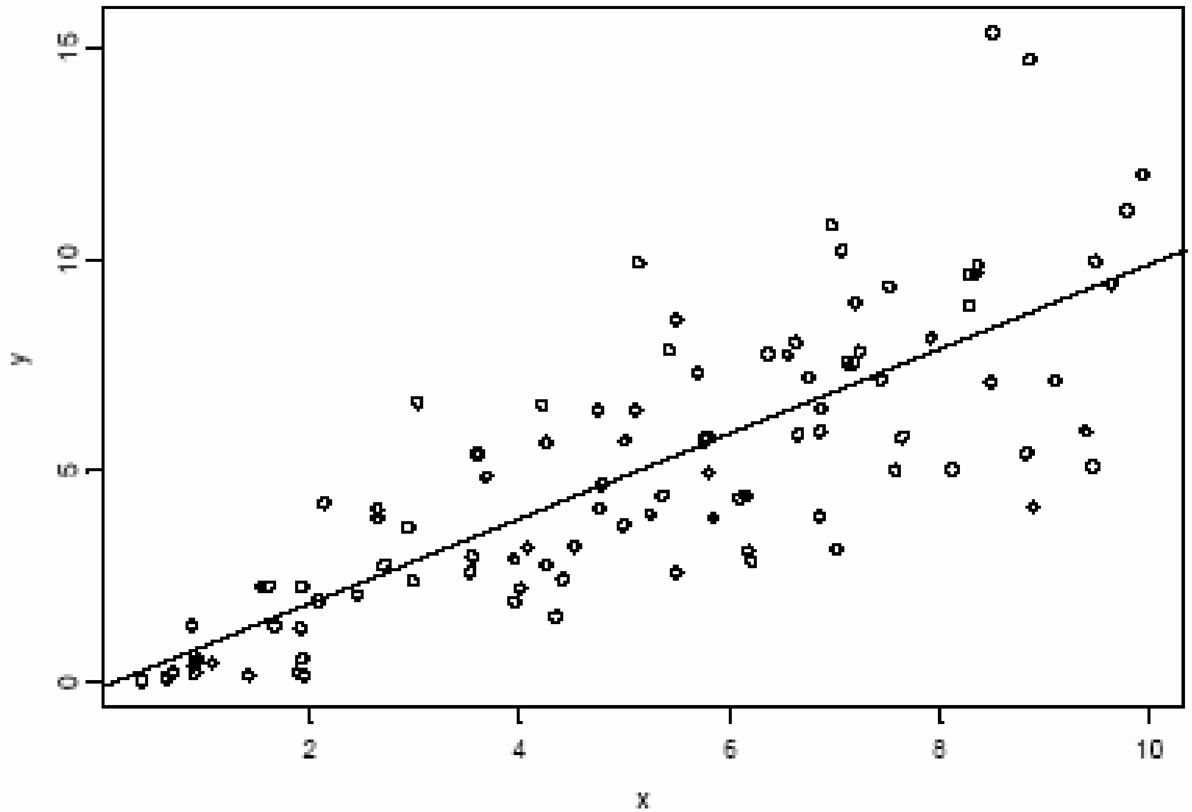
Relation between Income and Spending violates homoskedasticity

As income increases so do the errors (vertical distance from the predicted line)

Consequences of non-constant variance

- If "constant variance" is violated, LS estimates are still unbiased but SEs, tests, Confidence Intervals, and Prediction Intervals are incorrect

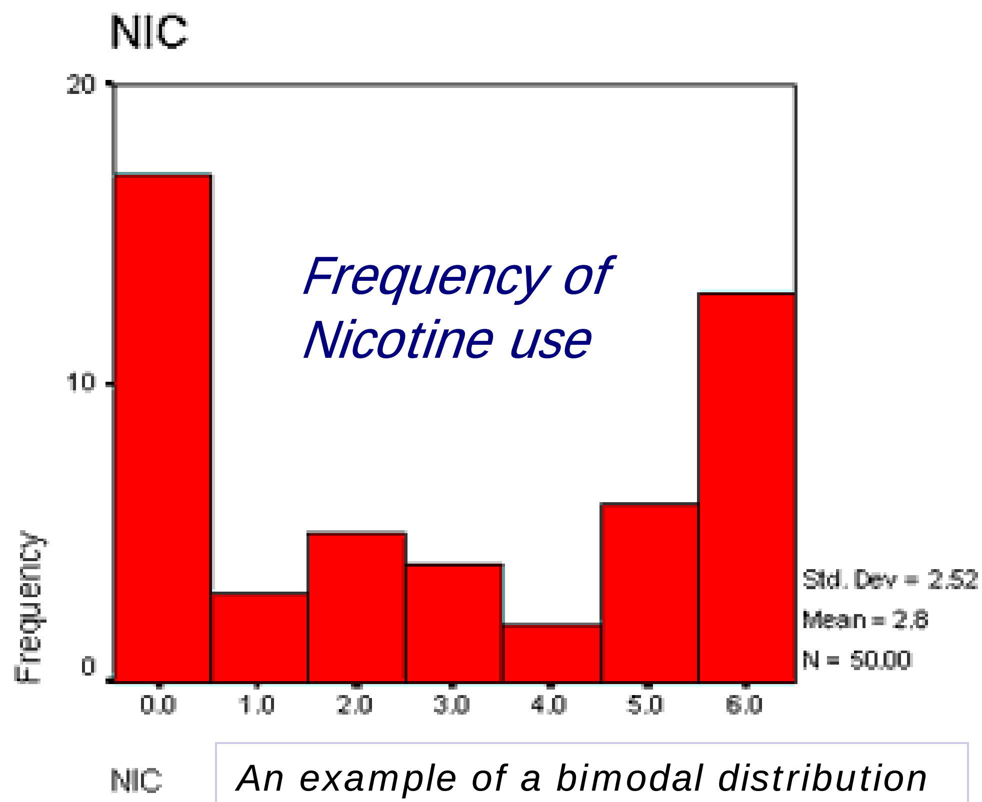
- However, the degree depends…



Violation of Normality

- Non-Normality

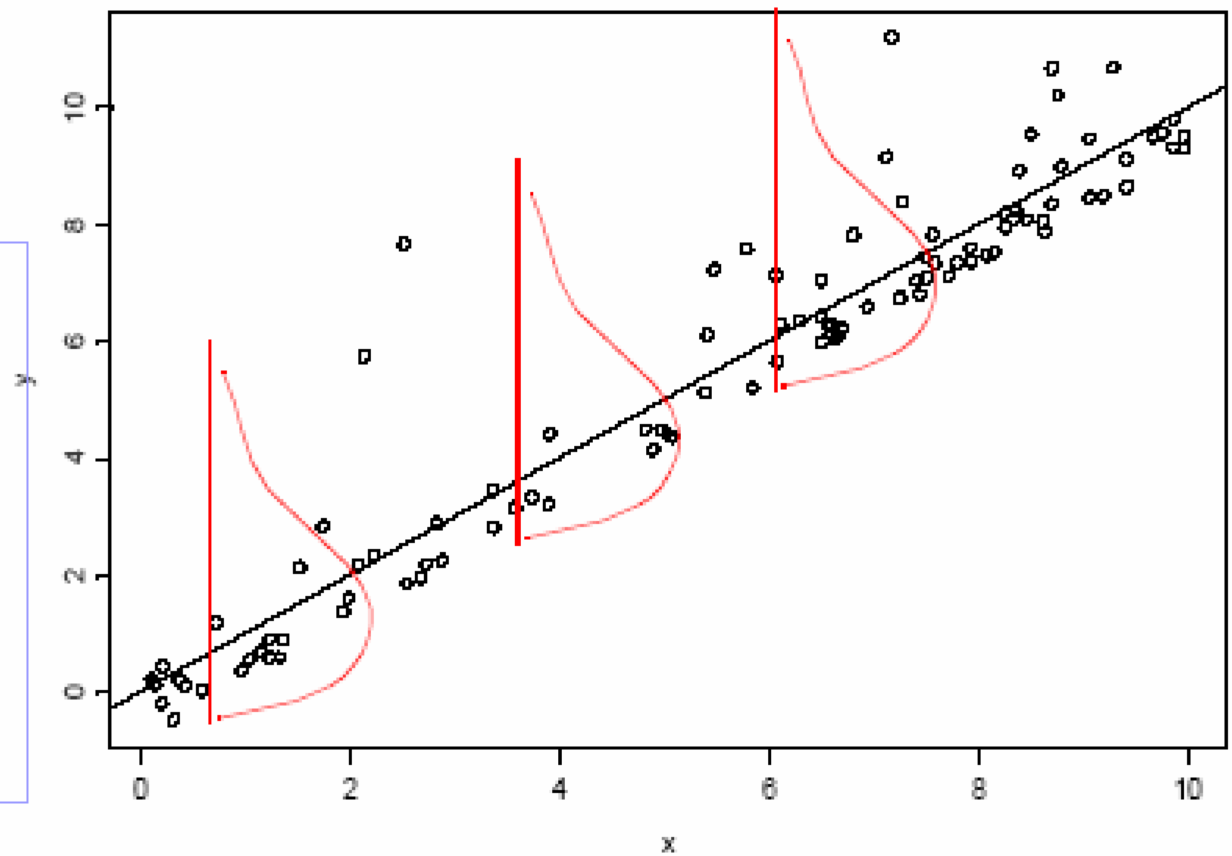
Nicotine use is characterized by a large number of people not smoking at all and another large number of people who smoke every day.



An example of a bimodal distribution

Consequence of non-Normality

- If "normality" is violated,
  - LS estimates are still unbiased
  - tests and CIs are quite robust
  - PIs are not

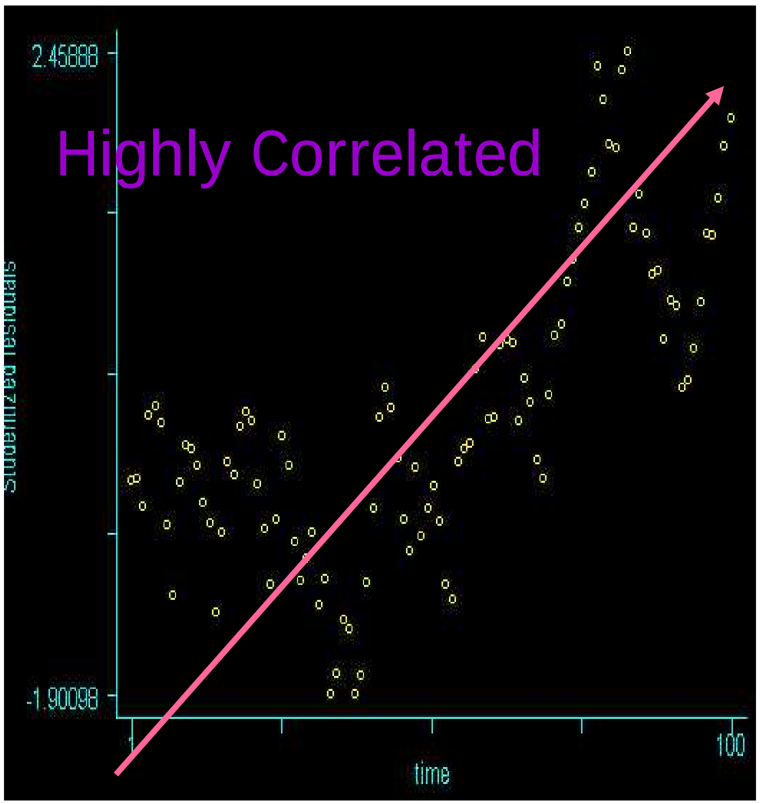


Of all the assumptions, this is the one that we need to be least worried about violating.

Why?

Violation of Non-independence

Residuals of GNP and Consumption over Time

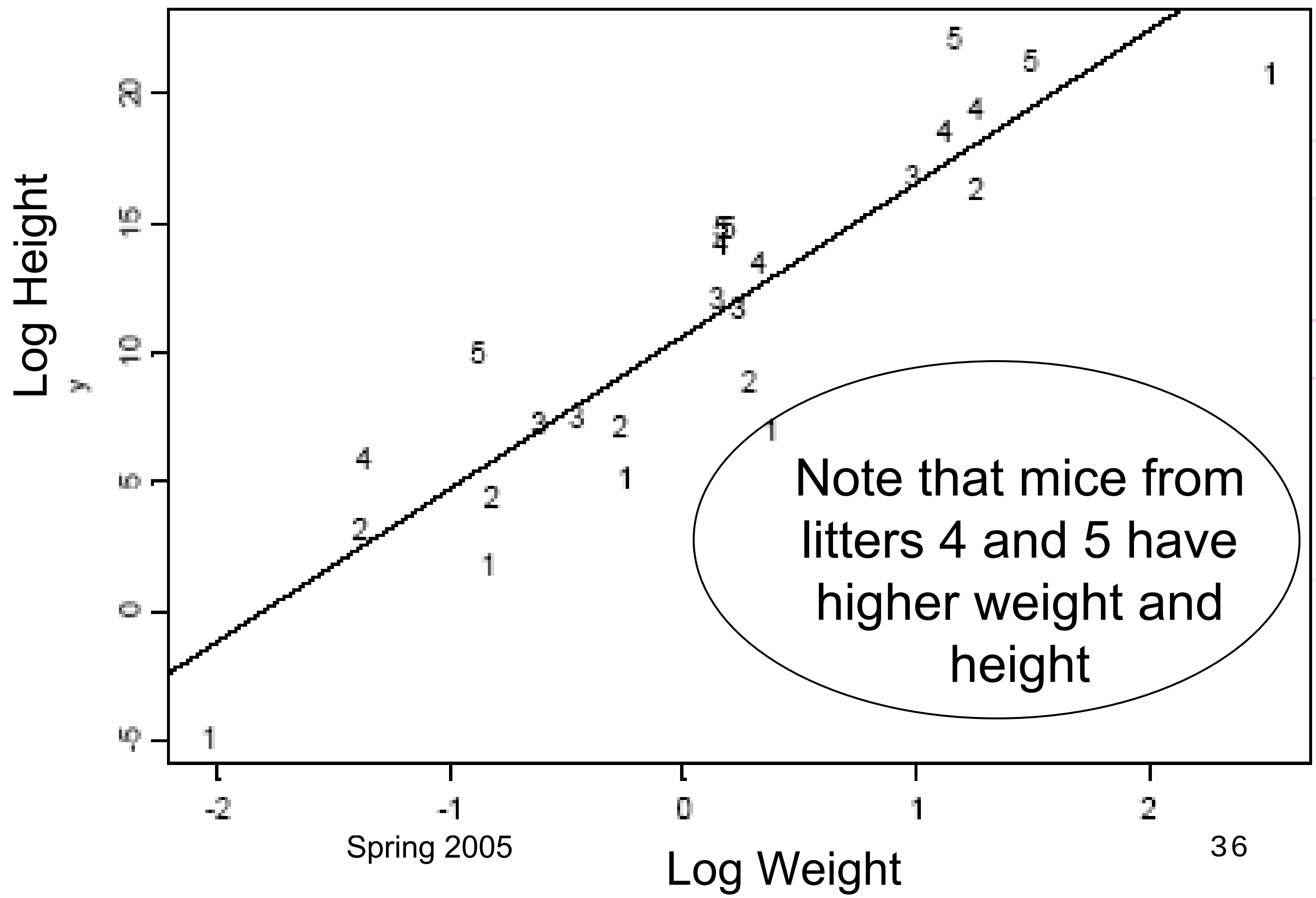


- Non-Independence
  - The independence assumption means that errors terms of two variables will not necessarily influence one another.
    - Technically, the RESIDUALS or error terms are uncorrelated.
  - The most common violation occurs with data that are collected over time or time series analysis.
    - Example: high tariff rates in one period are often associated with very high tariff rates in the next period.
    - Example: Nominal GNP and Consumption

Consequence of non-independence

- If "independence" is violated:
  - LS estimates are still unbiased
  - everything else can be misleading

Plotting code is litter (5 mice from each of 5 litters)



Robustness of least squares

- The "constant variance" assumption is important.
- Normality is not too important for confidence intervals and p-values, but is important for prediction intervals.
- Long-tailed distributions and/or outliers can heavily influence the results.
- Non-independence problems: serial correlation (Ch. 15) and cluster effects (we deal with this in Ch. 9-14).

Strategy for dealing with these potential problems

- Plots; Residual plots; Consider outliers (more in Ch. 11)
- Log Transformations (Display 8.6)

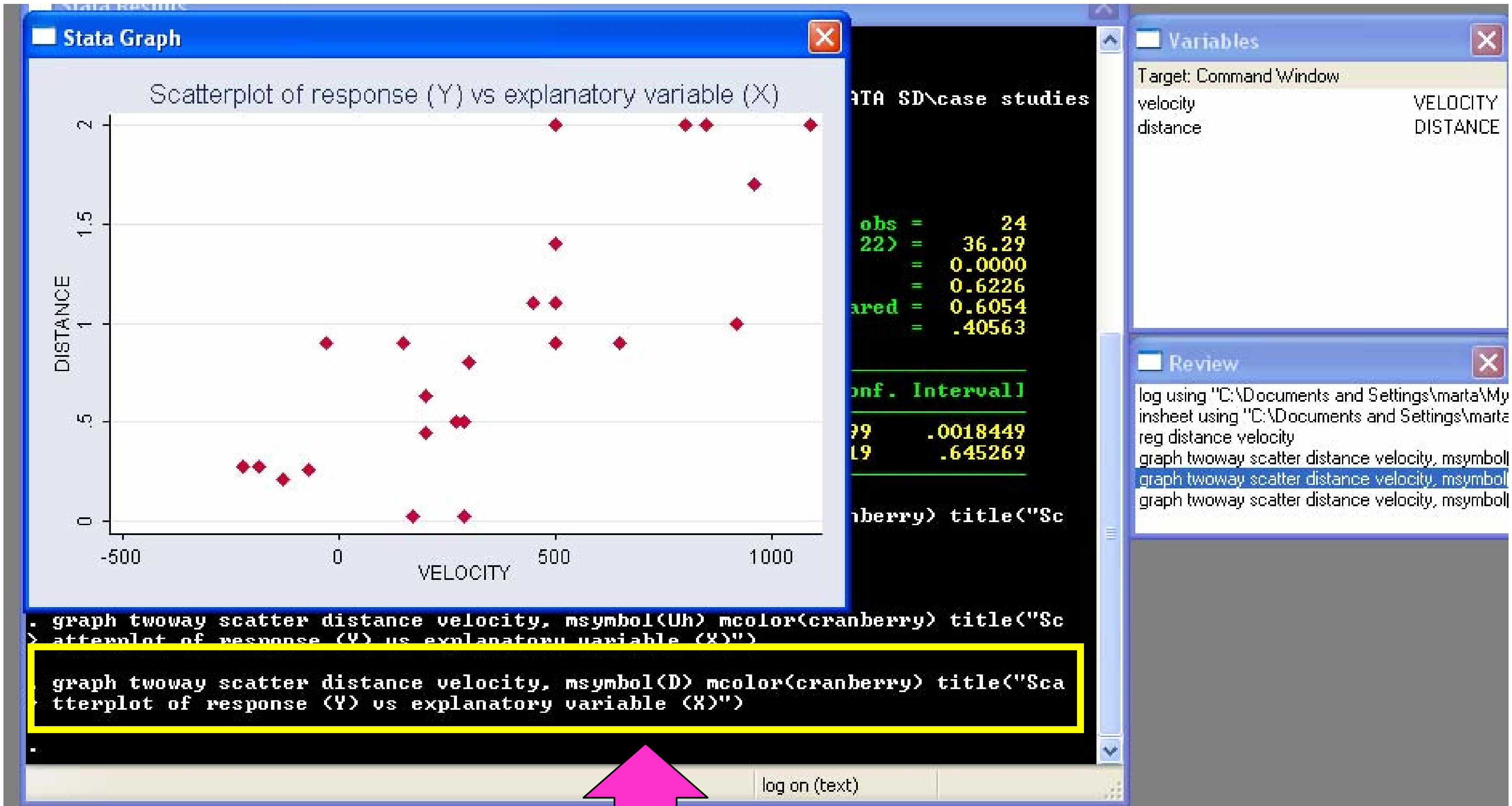
Tools for model checking

- Scatterplot of Y vs. X (see Display 8.6 p. 213)*
- Scatterplot of residuals vs. fitted values*

*Look for curvature, non-constant variance, and outliers

- Normal probability plot (p.224)
  - It is sometimes useful—for checking if the distribution is symmetric or normal (i.e. for PIs).
- Lack of fit F-test when there are replicates (Section 8.5).

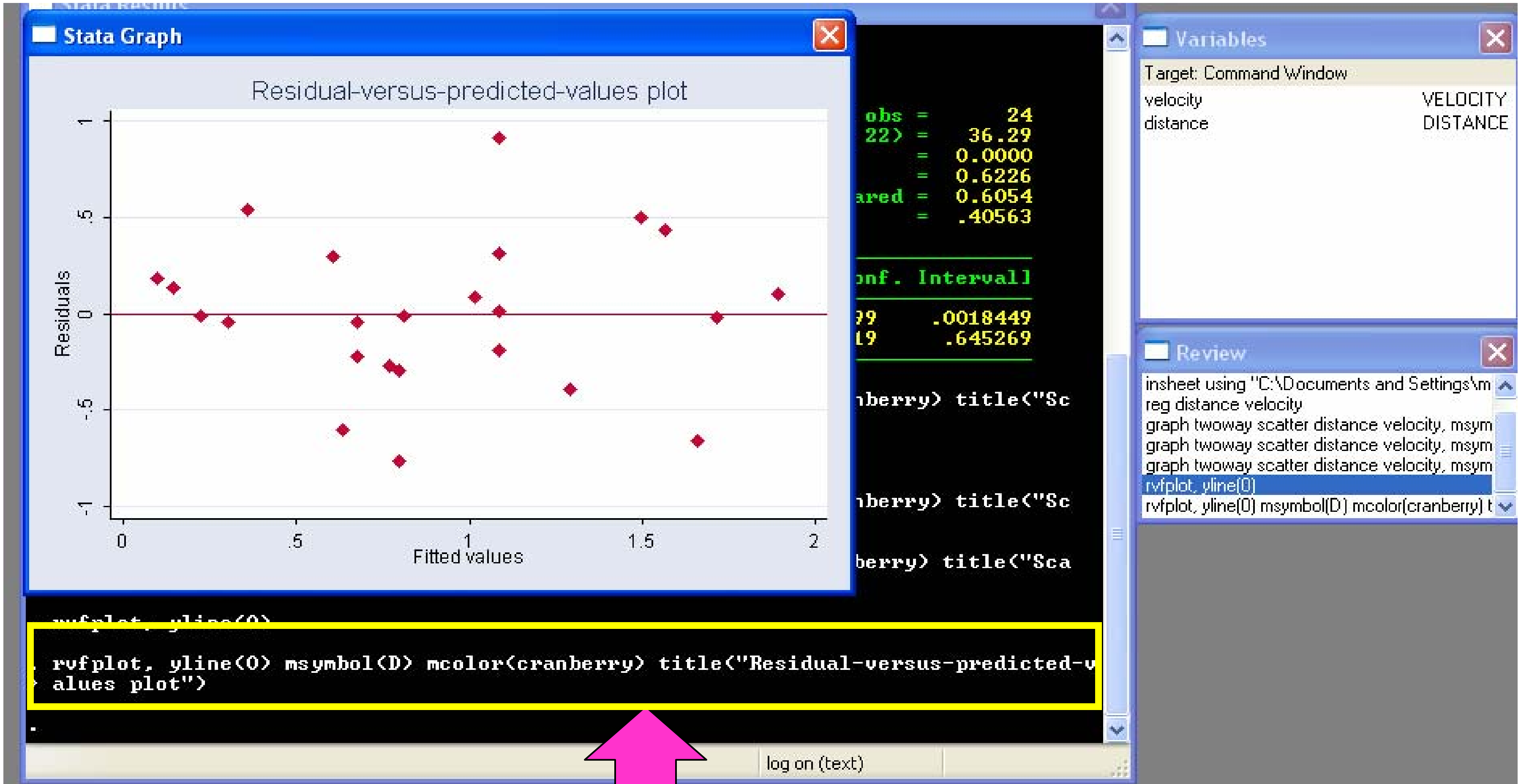
Scatterplot of Y vs. X



Command: **`graph twoway Y X`**

Case study: 7.01 page175

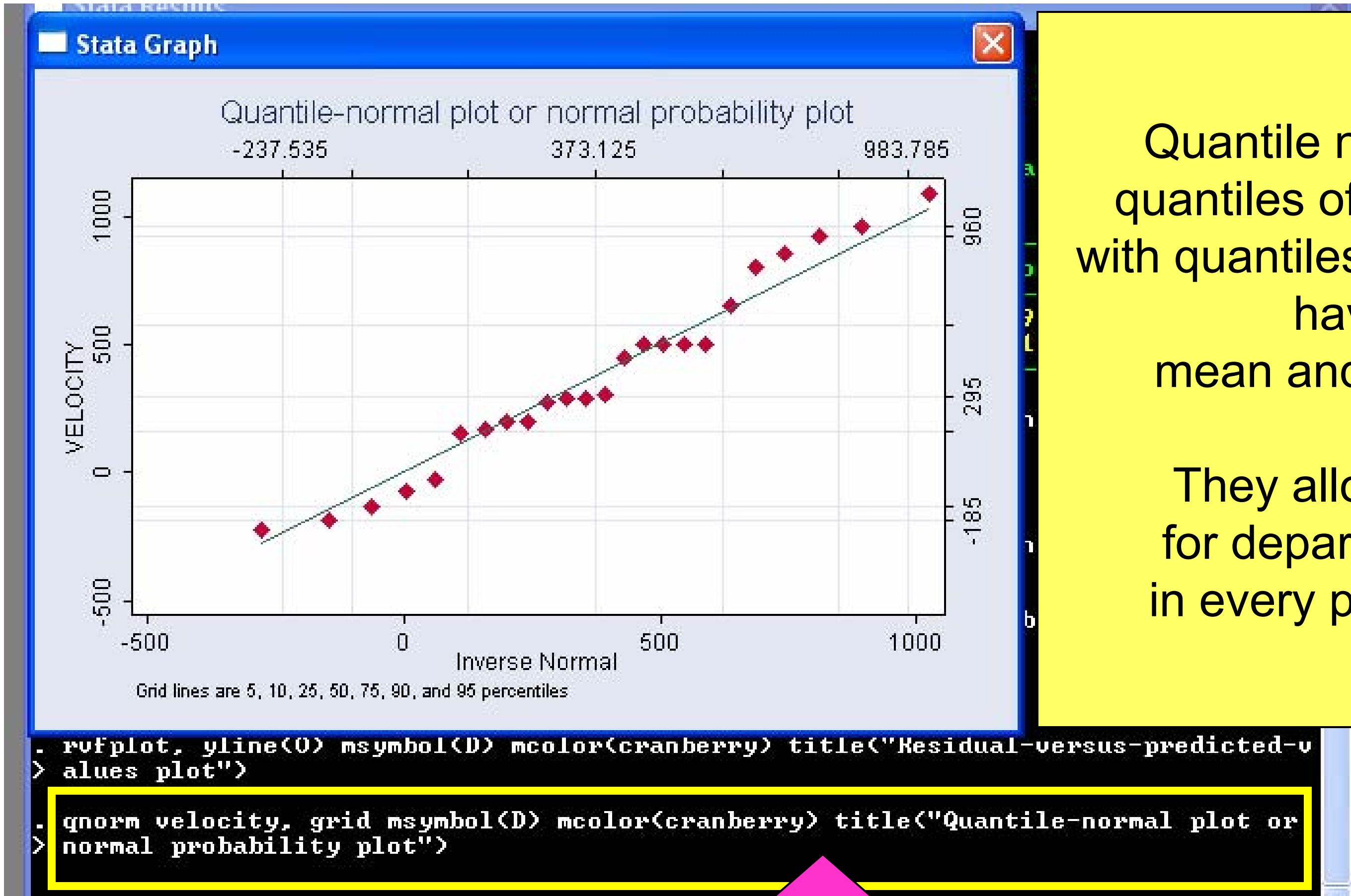
Scatterplot of residuals vs. fitted values



Command: **`rvfplot, yline(0)…`**

Case study: 7.01 page175

Normal probability plot (p.224)



Quantile normal plots compare quantiles of a variable distribution with quantiles of a normal distribution having the same mean and standard deviation.

They allow visual inspection for departures from normality in every part of the distribution.

Command: **`qnorm variable, grid`**

Case study: 7.01, page 175

Diagnostic plots of residuals

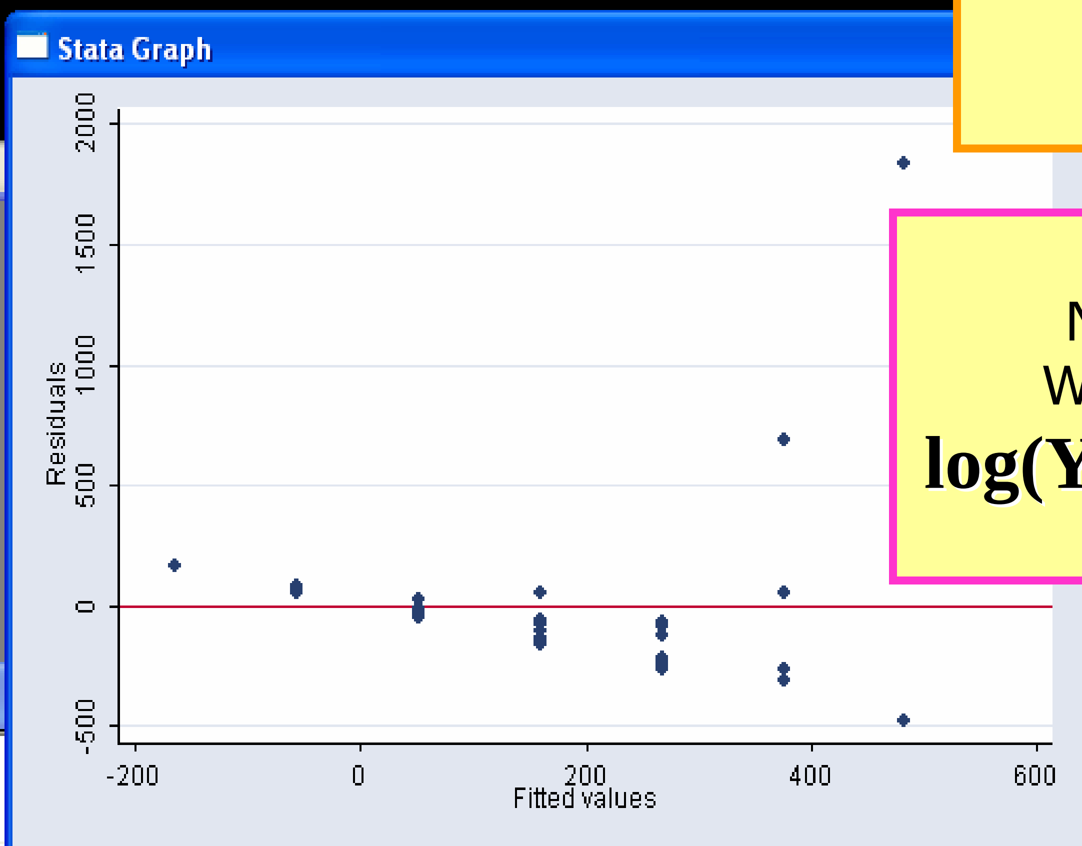
- Plot residuals versus fitted values almost always:
  - For simple reg. this is about the same as residuals vs. x
  - Look for outliers, curvature, increasing spread (funnel or horn shape); then take appropriate action.
- If data were collected over time, plot residuals versus time
  - Check for time trend and
  - Serial correlation
- If normality is important, use normal probability plot.
  - A straight line is expected if distribution is normal

Voltage Example (Case Study 8.1.2)

- Goal: to describe the distribution of breakdown time of an insulating fluid as a function of voltage applied to it.
  - Y=Breakdown time
  - X= Voltage
- Statistical illustrations
  - Recognizing the need for a log transformation of the response from the scatterplot and the residual plot
  - Checking the simple linear regression fit with a lack-of-fit F-test
  - Stata (follows)

Simple regression

```
. rvfplot, yline(0)

. reg time voltage

      Source |       SS       df       MS              Number of obs =      76
-------------+------------------------------           F(  1,    74) =   24.27
       Model |  2150408.26     1  2150408.26           Prob > F      =  0.0000
    Residual |  6557345.28    74  88612.774            R-squared     =  0.2470
-------------+------------------------------           Adj R-squared =  0.2368
       Total |  8707753.53    75  116103.38            Root MSE      =  297.68

------------------------------------------------------------------------------
        time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     voltage |  -53.95492   10.95264    -4.93   0.000    -75.77853   -32.13131
       _cons |   1886.169   364.4812     5.17   0.000     1159.925    2612.414
------------------------------------------------------------------------------

. rvfplot, yline(0)
```

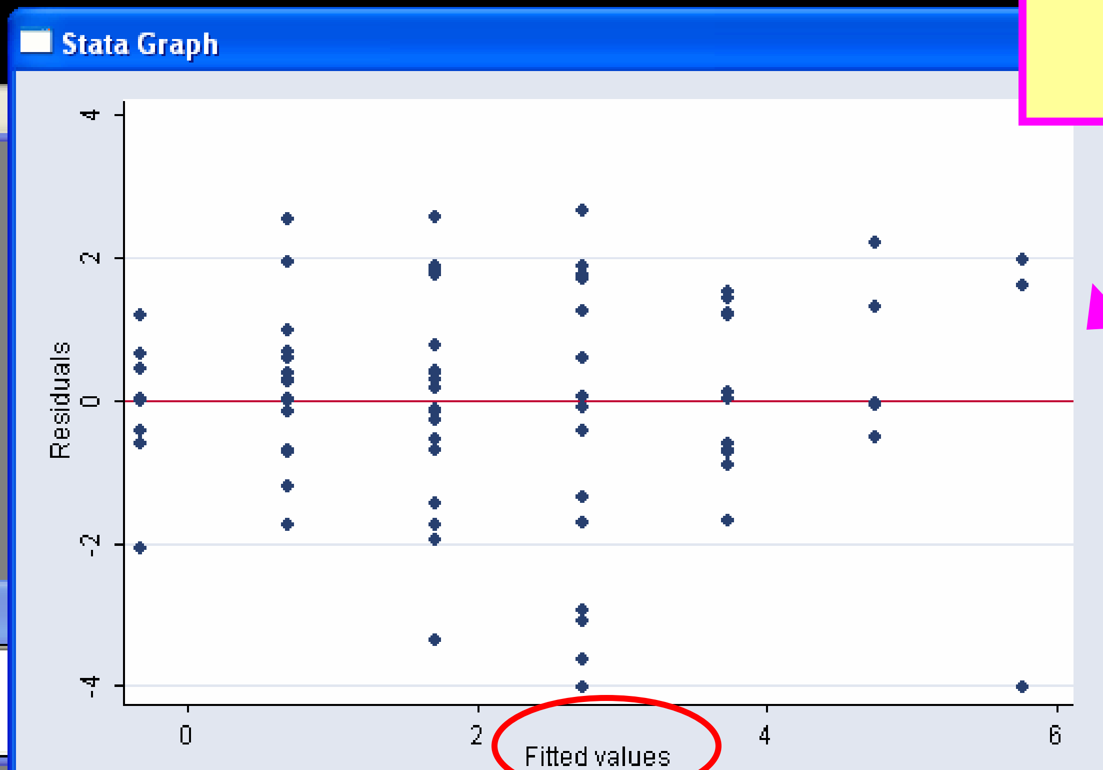


The residuals vs fitted values plot presents increasing spread with increasing fitted values

Next step:
We try with
log(Y) ~ log(time)

Simple regression with Y logged

```
(3 vars, 76 obs)

. gen ltime=log(time)

. reg ltime voltage

      Source |       SS       df       MS              Number of obs =      76
-------------+------------------------------           F(  1,    74) =   78.14
       Model |  190.151492     1  190.151492           Prob > F      =  0.0000
    Residual |   180.07484    74  2.43344378           R-squared     =  0.5136
-------------+------------------------------           Adj R-squared =  0.5070
       Total |  370.226332    75  4.93635109           Root MSE      =  1.5599

------------------------------------------------------------------------------
       ltime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     voltage |  -.5073649    .057396    -8.84   0.000    -.6217289    -.393001
       _cons |   18.95546   1.910019     9.92   0.000     15.14966    22.76125
------------------------------------------------------------------------------

. rvfplot, yline(0)
```



The residuals vs fitted values plot does not present any obvious curvature or trend in spread.

Interpretation after log transformations

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	Y	X	$\Delta y=\beta_1\Delta x$
Level-log	Y	log(X)	$\Delta y=(\beta_1/100)\%\Delta x$
Log-level	log(Y)	X	$\%\Delta y=(100\beta_1)\Delta x$
Log-log	log(Y)	log(X)	$\%\Delta y=(\beta_1)\%\Delta x$

Dependent variable logged

- $\mu\{log(Y)|X\} = \beta_0 + \beta_1 X$ is the same as:

(if the distribution of *log(Y)*, given *X*, is symmetric)

$$Median\ \{Y \,||\, X\} = e^{\beta_0 + \beta_1 X}$$

- As *X* increases by 1, what happens?

$$\frac{Median\{Y \mid X = x+1\}}{Median\{Y \mid X = x\}} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

$$Median\{Y \mid X = x+1\} = e^{\beta_1} Median\{Y \mid X = x\}$$

Interpretation of Y logged

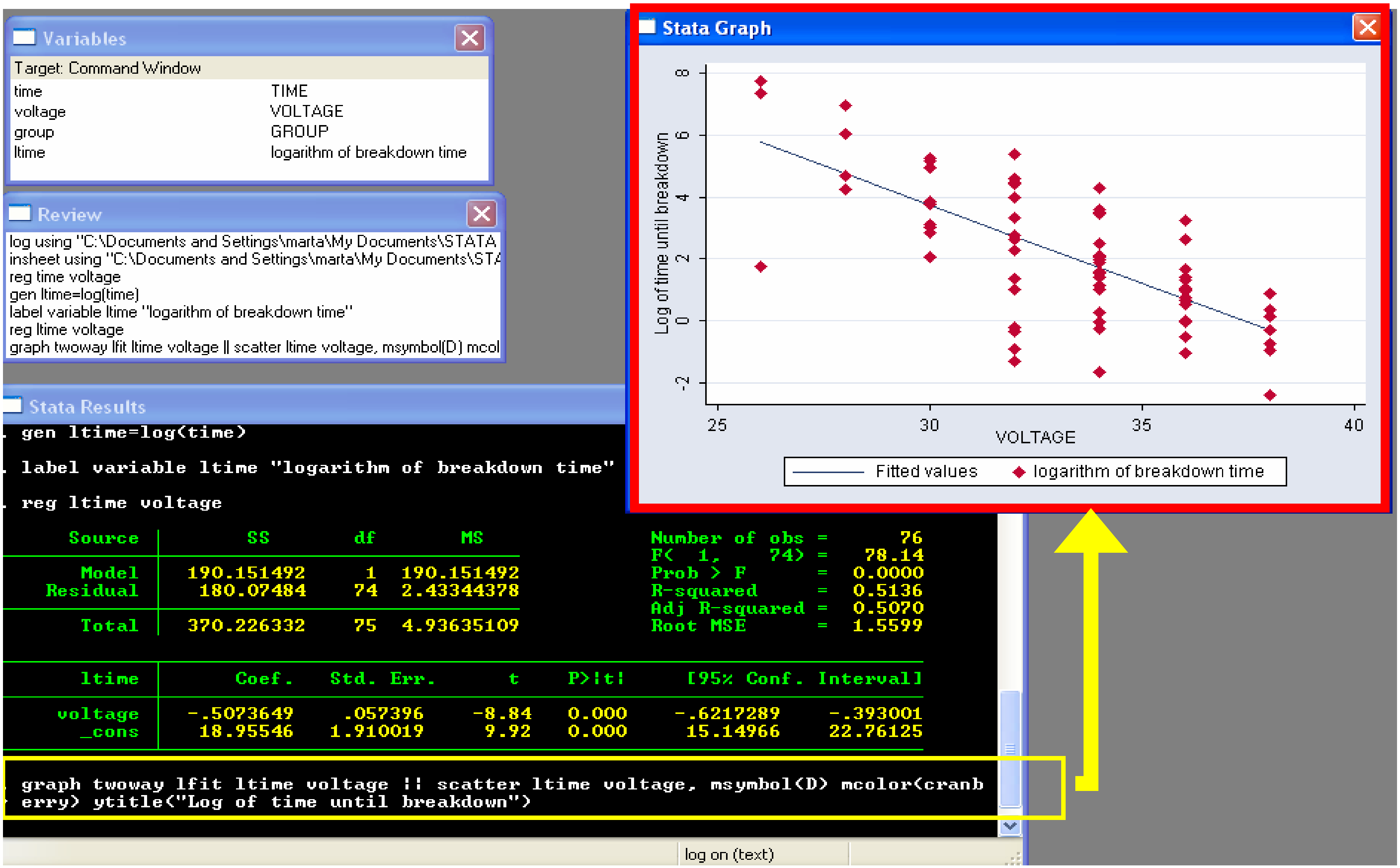
- "As *X* increases by 1, the median of *Y* changes by the multiplicative factor of e^{β_1}."

- Or, better:
  - If $\beta_1 > 0$: "As *X* increases by 1, the median of *Y* increases by $(e^{\beta_1} - 1) * 100\%$ "

- If $\beta_1 < 0$: "As *X* increases by 1, the median of *Y* decreases by $(1 - e^{\beta_1}) * 100\%$ "

Example: $\mu\{log(time)|voltage\} = \beta_0 - \beta_1\ voltage$

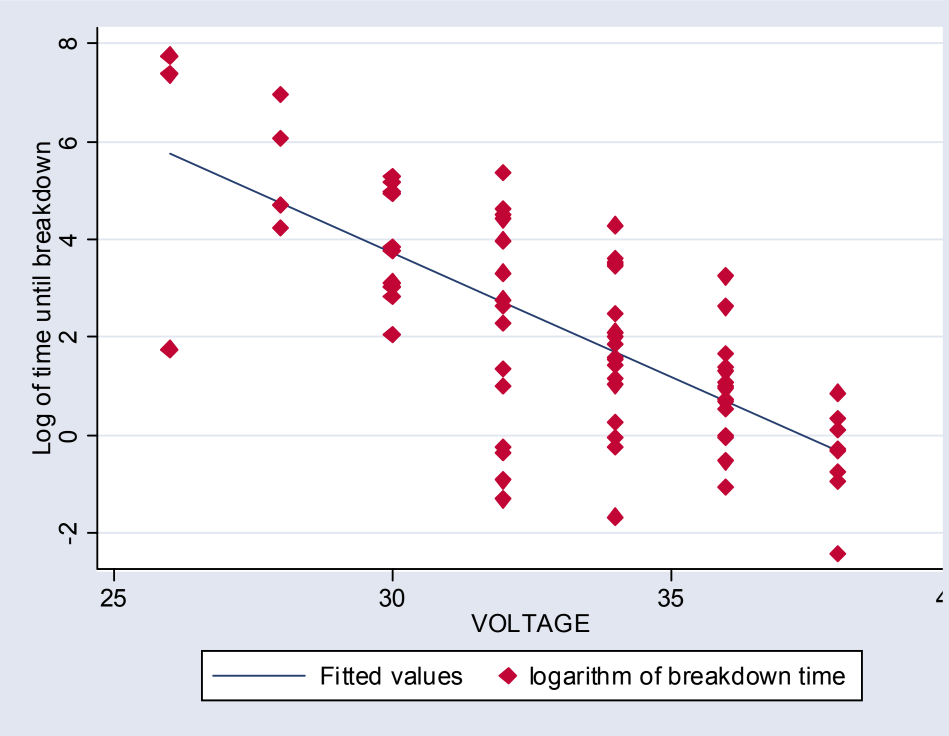
$1 - e^{-0.5} = .4$

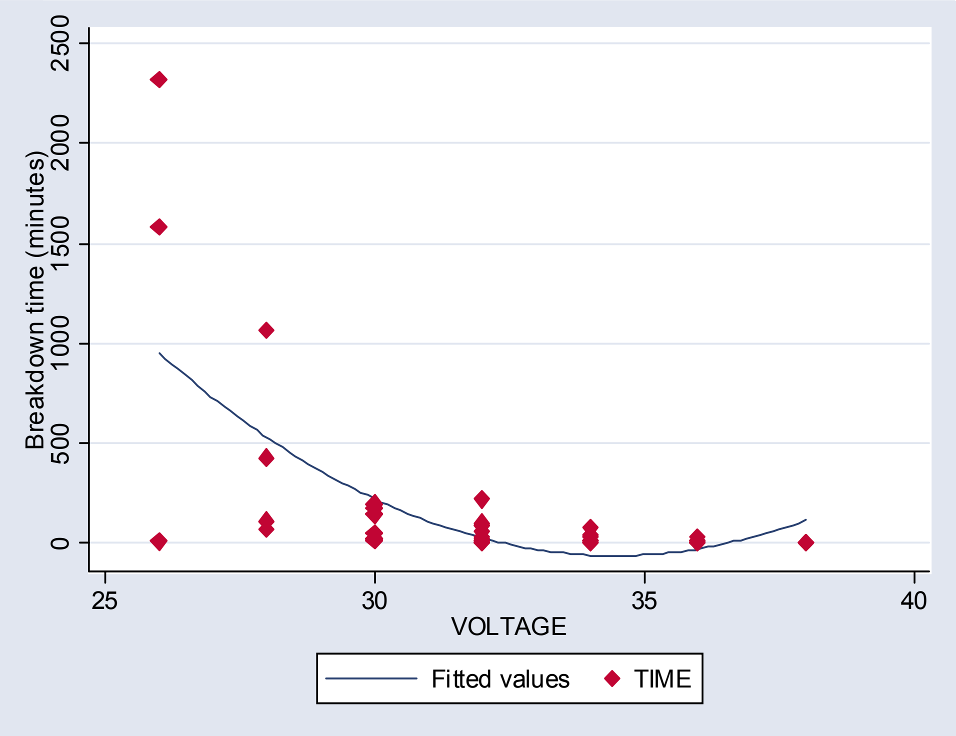


$$\mu\{log(time)|voltage\} = 18.96 - .507voltage$$

$$1- e^{-0.5}=.4$$

It is estimated that the median breakdown time decreases by 40% with each 1kV increase in voltage





If the explanatory variable (X) is logged

- If $\mu\{Y|log(X)\} = \beta_0 + \beta_1 log(X)$ then:
  - "Associated with each two-fold increase (i.e doubling) of *X* is a $\beta_1 log(2)$ change in the mean of *Y*."

- An example will follow:

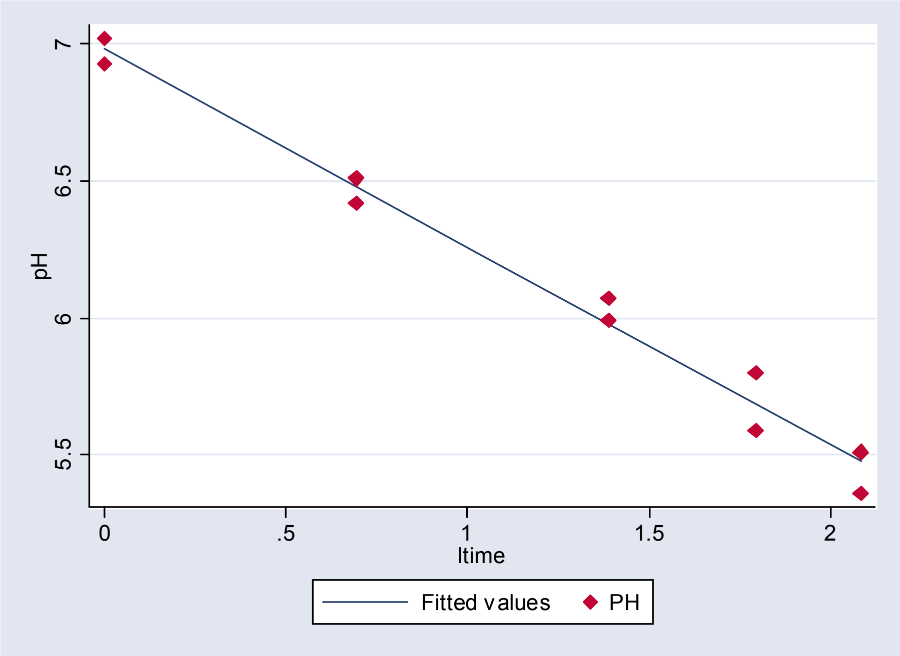
Example with X logged (Display 7.3 – Case 7.1):

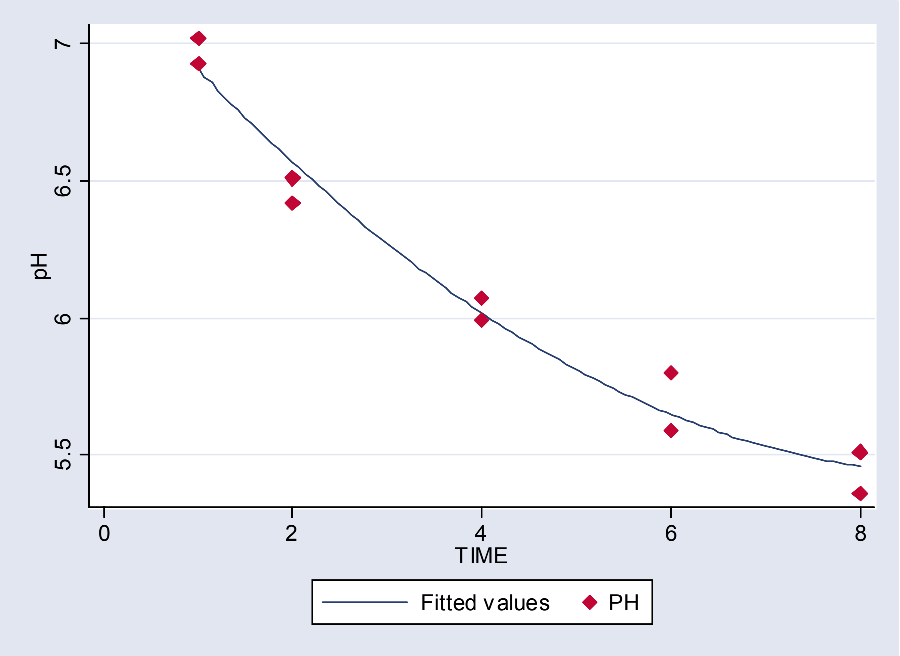
Y = pH

X = time after slaughter (hrs.)

estimated model: $\mu\{Y|log(X)\} = 6.98 - .73log(X)$.

$-.73'log(2) = -.5$ ➔ "It is estimated that for each doubling of time after slaughter (between 0 and 8 hours) the mean pH decreases by .5."





Both Y and X logged

- $\mu\{log(Y)|\log(X)\} = \beta_0 + \beta_1 log(X)$ is the same as:

- As *X* increases by 1, what happens?

If β_1>0: "As *X* increases by 1, the median of *Y* increases by $(e^{\log(2)\beta_1} - 1) * 100\%$ "

If β_1 < 0: "As *X* increases by 1, the median of *Y* decreases by $(1 - e^{\log(2)\beta_1}) * 100\%$ "

Example with Y and X logged Display 8.1 page 207

Y: number of species on an island
X: island area

$$\mu\{log(Y)|log(X)\} = \beta_0 - \beta_1\, log(X)$$

Stata Results

```
     species |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .0021112   .0004499     4.69   0.005     .0009548    .0032677
       _cons |   24.04928   9.074024     2.65   0.045     .7237545     47.3748
------------------------------------------------------------------------------

. gen lspecies=log(species)

. gen larea=log(area)

. reg lspecies larea

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  1,     5) =  425.30
       Model |  6.99619059     1  6.99619059           Prob > F      =  0.0000
    Residual |  .082249514     5  .016449903           R-squared     =  0.9884
-------------+------------------------------           Adj R-squared =  0.9861
       Total |   7.0784401     6  1.17974002           Root MSE      =  .12826

------------------------------------------------------------------------------
    lspecies |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .2496799   .0121069    20.62   0.000      .218558    .2808018
       _cons |   1.936508   .0881314    21.97   0.000     1.709959    2.163057
------------------------------------------------------------------------------
```

Y and X logged

$$\mu\{log(Y)|log(X)\} = 1.94 - .25\ log(X)$$

Since $e^{.25log(2)} = .19$

"Associated with each doubling of island area is a 19% increase in the median number of bird species"

Example: Log-Log

In order to graph the Log-log plot we need to generate two new variables (natural logarithms)