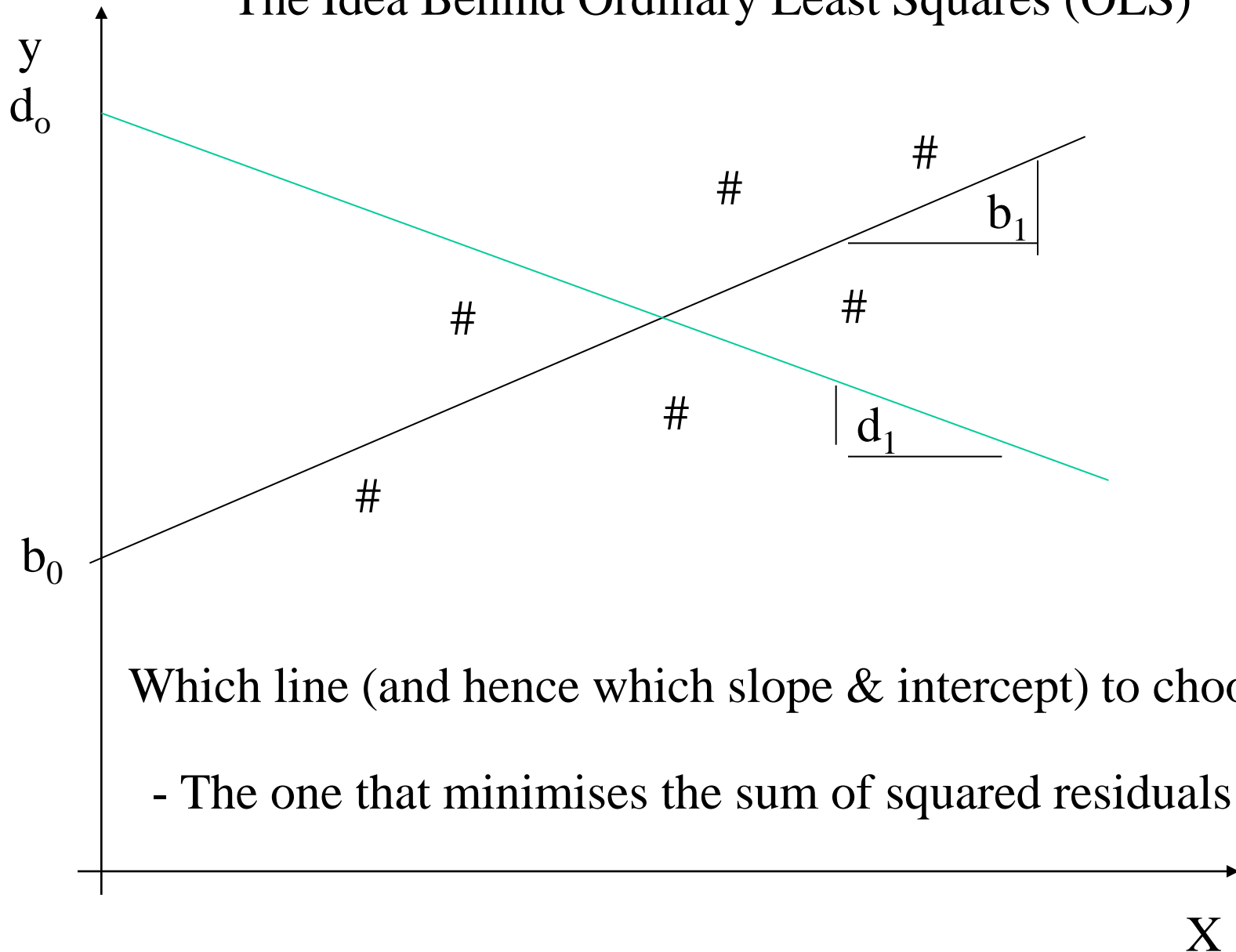


Things to do in Lecture 2

- OLS implies “minimising the sum of squared residuals”
- Derive formally how to estimate the value of the coefficients we are interested using OLS
- Run an OLS regression
- Interpret the regression output
- Measure how well the model fits the data

The Idea Behind Ordinary Least Squares (OLS)



We can then compare this predicted value with the actual value of the dependent variable and the difference between the actual and predicted value gives the **residual**

$$u_i = y_i - \hat{y}_i$$

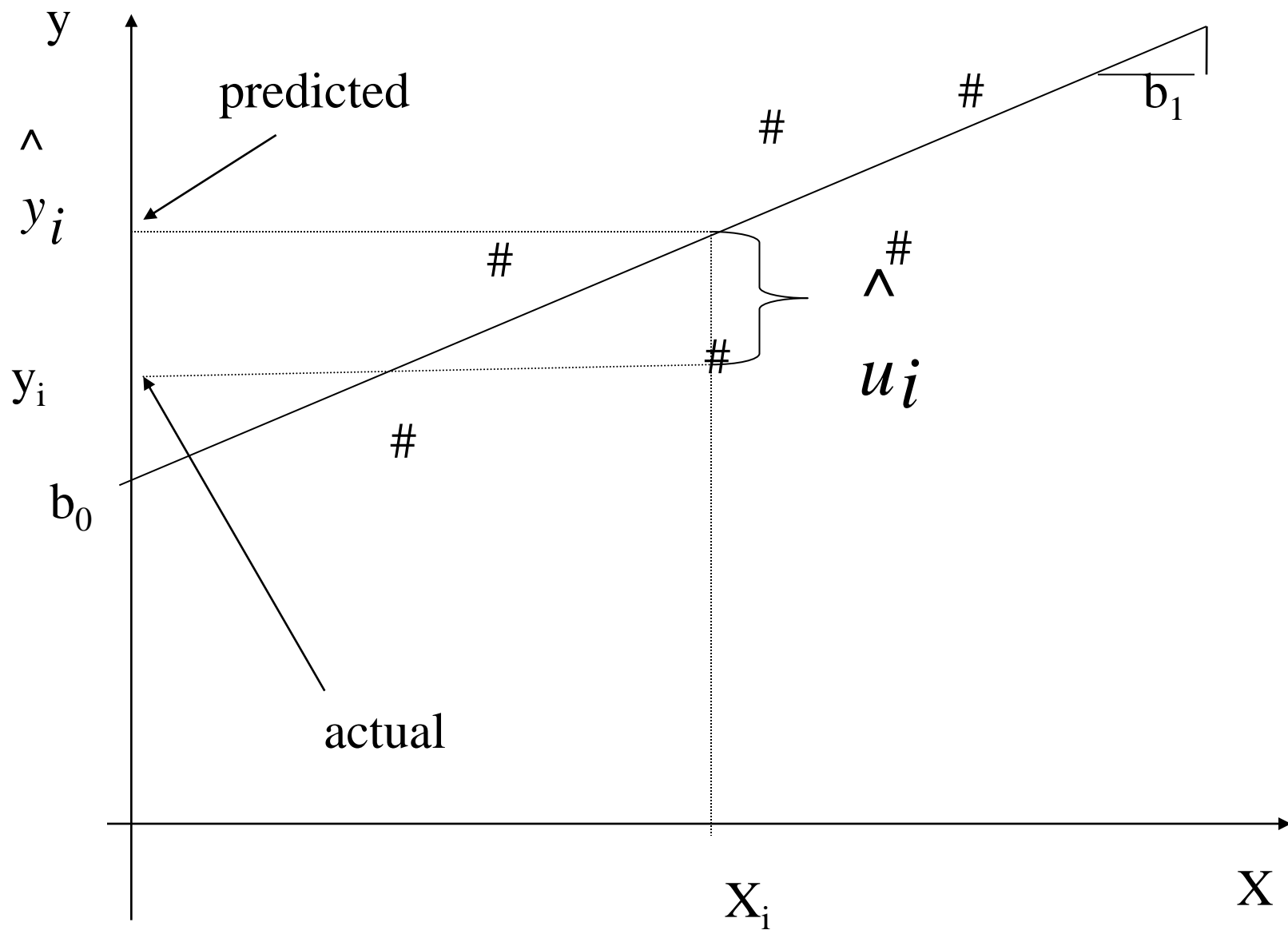
Which since

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

gives

$$u_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

The difference between the actual and predicted value is the residual



We can then compare this predicted value with the actual value of the dependent variable and the difference between the actual and predicted value gives the residual

$$u_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 X_i$$

(where the i subscript refers to the i^{th} individual or firm or time period in the data set)

Things to know about residuals

$$u_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 X_i$$

1. The larger the (absolute) value of the residual the worse the prediction

So intuitively then the line of best fit should be the one that delivers the smallest residual values for the each observation in the data set

2. Since the difference between the actual and predicted value gives the residual

$$u_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 X_i$$

a positive residual means

$$\hat{u}_i = y_i - \hat{y}_i > 0 \quad \text{so the model underpredicts}$$

and similarly

$$\hat{u}_i = y_i - \hat{y}_i < 0 \quad \text{The model overpredicts}$$

(larger than actual)

Given this..

Suppose we tried to minimise the sum of all the residuals in the data in an attempt to get the line of best fit

$$\sum_{i=1}^{N^{\wedge}} u_i$$

Whilst this might seem intuitive it will not work because it is possible that any **positive** residual will be offset by a **negative** residual in the summation and so the sum could be close to zero even if the overall fit of the regression were poor

We can avoid this problem is use instead the principle of **Ordinary Least Squares (OLS)**

Rather than minimise the sum of residuals, minimise the sum **of squared** residuals

$$\sum_{i=1}^N u_i^2$$

- Squaring ensures that values are always positive and so can never cancel each other out

-Also gives more “weight” to larger residuals and so harder to get away with a poor fit, (the larger any one residual in absolute value, the larger is the sum of squared residuals (RSS))

Consider the following simple example

$N=2$ and want to fit a straight line $y=b_0 + b_1X$
 thru' the following data points using principle of OLS
 (min sum of squared residuals)

$$(Y_1=3 \quad X_1 =1)$$

$$(Y_2=5 \quad X_2=2)$$

It follows that can write estimated residual for the 1st observation

$$u_1 = y_1 - \hat{y}_1$$

using $\hat{y}_1 = b_0 + b_1(X_1)$

$$\hat{y}_1 = b_0 + b_1(1)$$

and $y_1=3$ so

$$u_1 = 3 - b_0 - b_1(1)$$

and similarly for the 2nd observation

$$u_2 = y_2 - \hat{y}_2 = 5 - b_0 - b_1(2)$$

OLS: minimise the sum of squared residuals

$$S = u_1^2 + u_2^2$$

$$S = (3 - \hat{b}_0 - \hat{b}_1)^2 + (5 - \hat{b}_0 - 2\hat{b}_1)^2$$

Expanding the terms in brackets

$$S = (9 + \hat{b}_0^2 + \hat{b}_1^2 - 6\hat{b}_0 - 6\hat{b}_1 + 2\hat{b}_0\hat{b}_1) \\ + (25 + \hat{b}_0^2 + 4\hat{b}_1^2 - 10\hat{b}_0 - 20\hat{b}_1 + 4\hat{b}_0\hat{b}_1)$$

Adding together like terms

$$S = (2\hat{b}_0^2 + 5\hat{b}_1^2 - 16\hat{b}_0 - 26\hat{b}_1 + 6\hat{b}_0\hat{b}_1 + 25) \quad (\text{A})$$

Now need to find values of \hat{b}_0 and \hat{b}_1

which minimise this sum, $S = (2\hat{b}_0^2 + 5\hat{b}_1^2 - 16\hat{b}_0 - 26\hat{b}_1 + 6\hat{b}_0\hat{b}_1 + 25)$

Using the rules of calculus we know the first order condition for minimisation are:

$$\frac{dS}{d\hat{b}_0} = 0$$

$$\frac{dS}{d\hat{b}_1} = 0$$

→

$$4\hat{b}_0 + 6\hat{b}_1 - 16 = 0$$

$$6\hat{b}_0 + 10\hat{b}_1 - 26 = 0$$

This gives 2 simultaneous equations

$$2\hat{b}_0 + 3\hat{b}_1 = 8$$

$$3\hat{b}_0 + 5\hat{b}_1 = 13$$

which can solve for unknown values of \hat{b}_0 and \hat{b}_1

using rules for simultaneous equations

$$\hat{b}_0 = 1 \quad \hat{b}_1 = 2$$

So the estimated regression line becomes $\hat{Y} = 1 + 2X$

ie the intercept (constant) with the y axis is at 1 and the slope of the straight line is 2

Basic idea underlying OLS is to choose a “line of best fit”

-Choose a straight line that passes through the data and minimises the sum of squared residuals

Now need to do this more generally so can apply the technique to any possible combination of (x, y) data pairs and any number of observations

If we wish to fit a (straight) line through N (rather than 2) observations, then the OLS principle is still the same ie choose

\hat{b}_0 and \hat{b}_1 to minimise

$$S = u_1^2 + u_2^2 + \dots + u_N^2 = \sum_{i=1}^N u_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

(where now the summation runs from 1 to N rather than 1 to 2)

$$\text{sub. in } \hat{y}_i = \hat{b}_0 - \hat{b}_1 X_i$$

$$\begin{aligned} S &= (Y_1 - \hat{b}_0 - \hat{b}_1 X_1)^2 + \dots + (Y_n - \hat{b}_0 - \hat{b}_1 X_n)^2 \\ &= Y_1^2 + \hat{b}_0^2 + \hat{b}_1^2 X_1^2 - 2\hat{b}_0 Y_1 - 2\hat{b}_1 X_1 Y_1 + 2\hat{b}_0 \hat{b}_1 X_1 \\ &\quad + \dots \\ &\quad + Y_n^2 + \hat{b}_0^2 + \hat{b}_1^2 X_n^2 - 2\hat{b}_0 Y_n - 2\hat{b}_1 X_n Y_n + 2\hat{b}_0 \hat{b}_1 X_n \\ &= \sum Y_i^2 + N \hat{b}_0^2 + \hat{b}_1^2 \sum X_i^2 - 2\hat{b}_0 \sum Y_i - 2\hat{b}_1 \sum X_i Y_i + 2\hat{b}_0 \hat{b}_1 \sum X_i \end{aligned}$$

This is just a generalised version of (A) above

Again, find values of \hat{b}_0 and \hat{b}_1

which minimise this sum, using the same simple calculus rules

$$\frac{dS}{d\hat{b}_0} = 0 \quad \text{and} \quad \frac{dS}{d\hat{b}_1} = 0$$

Now these two (1st order) minimisation conditions give

$$\frac{\partial S}{\partial b_0} = 0 \Rightarrow 2N\hat{b}_0 - 2\sum Y_i + 2\hat{b}_1 \sum X_i = 0 \quad (1)$$

$$\frac{\partial S}{\partial b_1} = 0 \Rightarrow 2\hat{b}_1 \sum X_i^2 - 2\sum X_i Y_i + 2\hat{b}_0 \sum X_i = 0 \quad (2)$$

and again we have 2 simultaneous equations (called the normal equations) which can again solve for

$$\hat{b}_0 \quad \text{and} \quad \hat{b}_1$$

Using the fact that the sample means of Y and X

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} \Leftrightarrow N \bar{Y} = \sum_{i=1}^N y_i \quad \bar{X} = \frac{\sum_{i=1}^N x_i}{N} \Leftrightarrow N \bar{X} = \sum_{i=1}^N x_i$$

can re-write (1)
$$2N \hat{b}_0 - 2 \sum Y_i + 2 \hat{b}_1 \sum X_i = 0$$

$$2N \hat{b}_0 - 2N \bar{Y} + 2 \hat{b}_1 N \bar{X} = 0$$

and so obtain the formula to calculate the OLS estimate of the intercept

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} \quad (***) \text{ learn this } (**)$$
(3)

Sub. $\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$ into (2)

$$2\hat{b}_1 \sum X_i^2 - 2\sum \sum X_i Y_i + 2\hat{b}_0 \sum X_i = 0$$

gives

$$\hat{b}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{b}_1 \bar{X}) \sum X_i = 0$$

and simplifying

$$\hat{b}_1 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{b}_1 \bar{X}) N \bar{X} = 0$$

collecting terms

$$\hat{b}_1 \left(\sum X_i^2 - N \bar{X}^2 \right) = \sum X_i Y_i - N \bar{X} \bar{Y}$$

Dividing both sides by 1/N

$$\hat{b}_1 \left(\frac{1}{N} \sum X_i^2 - \bar{X}^2 \right) = \frac{1}{N} \sum X_i Y_i - \bar{X} \bar{Y}$$

$$\hat{b}_1 \left(\frac{1}{N} \sum (X_i - \bar{X})^2 \right) = \frac{1}{N} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

which gives the formula to calculate the OLS estimate of the slope

$$\hat{b}_1 \text{Var}(X) = \text{Cov}(X, Y)$$

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

(**** learn this ****)

So $\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$ $\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$

are how the computer determines the size of the intercept and the slope respectively in an OLS regression

The OLS equations give a nice, clear intuitive meaning about the influence of the variable X on the size of the slope, since it shows that:

- i) the greater the covariance between X and Y, the larger the (absolute value of) the slope
- ii) the smaller the variance of X, the larger the (absolute value of) the slope

It is equally important to be able to interpret the effect of an estimated regression coefficient

Given OLS essentially passes a straight line through the data, then given

$$\hat{y} = \hat{b}_0 - \hat{b}_1 X$$

$$\frac{d \hat{y}}{d X} = \hat{b}_1$$

So the OLS estimate of the slope will give an *estimate* of the *unit change* in the dependent variable y following a *unit change* in the level of the explanatory variable

$$d \hat{y} = \hat{b}_1 d X$$

(so you need to be aware of the units of measurement of your variables in order to be able to interpret what the OLS coefficient is telling you)

Stata/SE 10.1 - C:\qm2\cons09.dta

File Edit Data Graphics Statistics User Window Help

Review

```

1 use "C:\qm2\cons09.dta", dea
2 reg cons income
3 predict chat
4 two (scatter cons income) (line

```

Results

Statistics/Data Analysis

StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)

Special Edition

Single-user Stata for windows perpetual license:
Serial number: 81910512684
Licensed to: Jonathan Wadsworth
Economics, Royal Holloway

Notes:

1. (/m# option or -set memory-) 10.00 MB allocated to data
2. (/v# option or -set maxvar-) 5000 maximum variables

```

. use "C:\qm2\cons09.dta", clear
. reg cons income

```

Source	SS	df	MS	Number of obs =
Model	4.4090e+12	1	4.4090e+12	62
Residual	1.1825e+10	60	197081123	F(1, 60) = 22371.65
Total	4.4209e+12	61	7.2473e+10	Prob > F = 0.0000

R-squared = 0.9973
Adj R-squared = 0.9973
Root MSE = 14039

	cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income		.8541788	.0057108	149.57	0.000	.8427555 .8656022
_cons		-4590.321	4525.946	-1.01	0.315	-13643.56 4462.919

```

. predict chat
(option xb assumed; fitted values)
. two (scatter cons income) (line chat income)

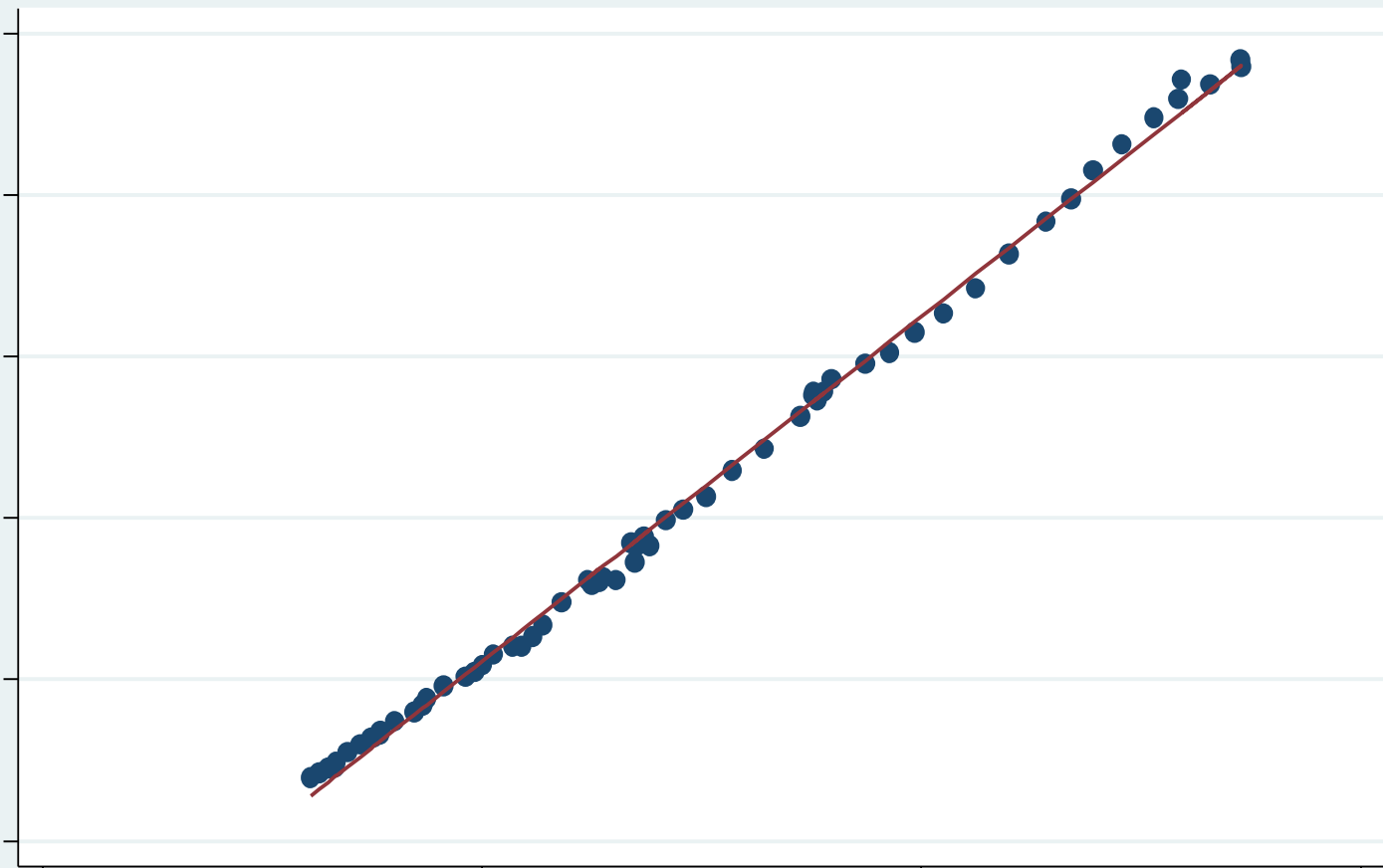
```

Command

C:\data

start

17:18 Thursday 13/01/2011



0 500000 1000000 1500000
Gross Domestic Product: chained volume measures: Seasonally adjusted

● Total Final Consumption Expenditure (NSA CVM) ABKX — Fitted values