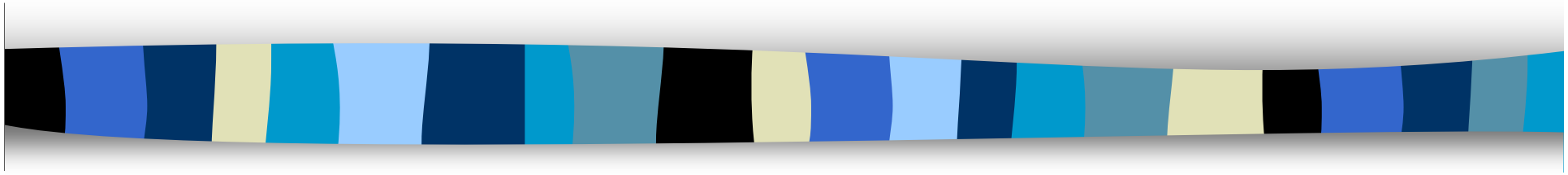
Graduate School

SSSII

Gwilym Pryce



Lecture 3: Misspecification: Non-linearities

Summary of Lecture 2:

- 1. ANOVA in regression
- 2. Prediction
- 3. F-Test
- 4. Regression assumptions
- 5. Properties of OLS estimates

TSS = REGSS + RSS

- The sum of squared deviations of *y* from the mean (i.e. the numerator in the variance of *y* equation) is called the
  TOTAL SUM OF SQUARES **(TSS)**
- The sum of squared deviations of error *e* is called the
  RESIDUAL SUM OF SQUARES* **(RSS)**
  * sometimes called the "error sum of squares"
- The difference between TSS & RSS is called the

  REGRESSION SUM OF SQUARES# **(REGSS)**
  #the REGSS is sometimes called the "explained sum of squares" or "model sum of squares"

  $\Rightarrow$ **TSS = REGSS + RSS**
- R^2 = **REGSS**/ **TSS**

4. Regression assumptions

For estimation of *a* and *b* and for regression inference to be correct:

- 1. Equation is correctly specified:
  - Linear in parameters (can still transform variables)
  - Contains all relevant variables
  - Contains no irrelevant variables
  - Contains no variables with measurement errors
- 2. Error Term has zero expected mean and zero conditional mean
  - $E(e_i) = 0$ Unobservable factors has zero mean
  - $E(e_i|x) = 0$ Error term is not dependent on (is unrelated to) x.
- 3. Error Term has constant variance

- 4. Error Term is not autocorrelated
  - I.e. correlated with error term from previous time periods
- 5. No linear relationship between RHS variables
  - I.e. no “multicolinearity”

5. Properties of OLS estimates

- If the above assumptions are met, OLS estimates are said to be BLUE:
  - Best — I.e. most efficient = least variance
  - Linear — I.e. best amongst linear estimates
  - Unbiased — I.e. in repeated samples, mean of $b = \beta$
  - Estimates — I.e. estimates of the population parameters.

Plan of Lecture 3:

- 1. Consequences of non-linearities
- 2. Testing for non-linearities
  - (a) visual inspection of plots
  - (b) t-statistics
  - (c) structural break tests
- 3. Solutions
  - (a) transform variables
  - (b) split the sample
  - (c) dummies
  - (d) use non-linear estimation techniques

Introduction

- OLS assumes linear relationships between y and each x.
- This rules out many possibilities that could plausibly occur in social science relationships
  - E.g. Relationship between income, education and years of experience…

Is the imposed linearity likely to be realistic? What sort of relationships are more likely?

– e.g. Income, education & experience

Model		Unstandardized Coefficients B	Std. Error
1	(Constant)	-4.200	23.951
	X1	1.450	1.789
	X2	2.633	3.117

a. Dependent Variable: Y

where x1 = post-school education,

x2 = experience

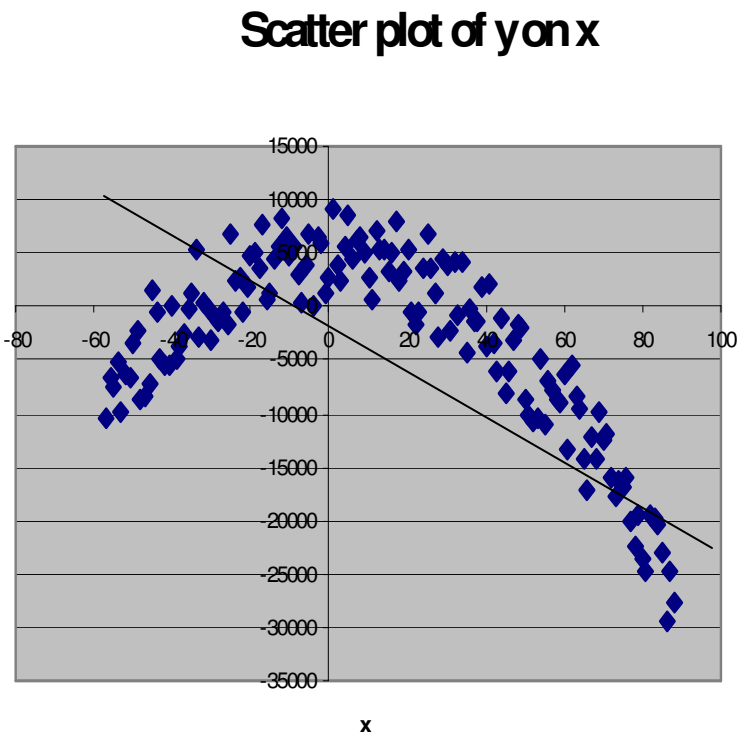
– Implies the following equation:

$$\hat{y} = -4.2 + 1.45\, x_1 + 2.63\, x_2$$

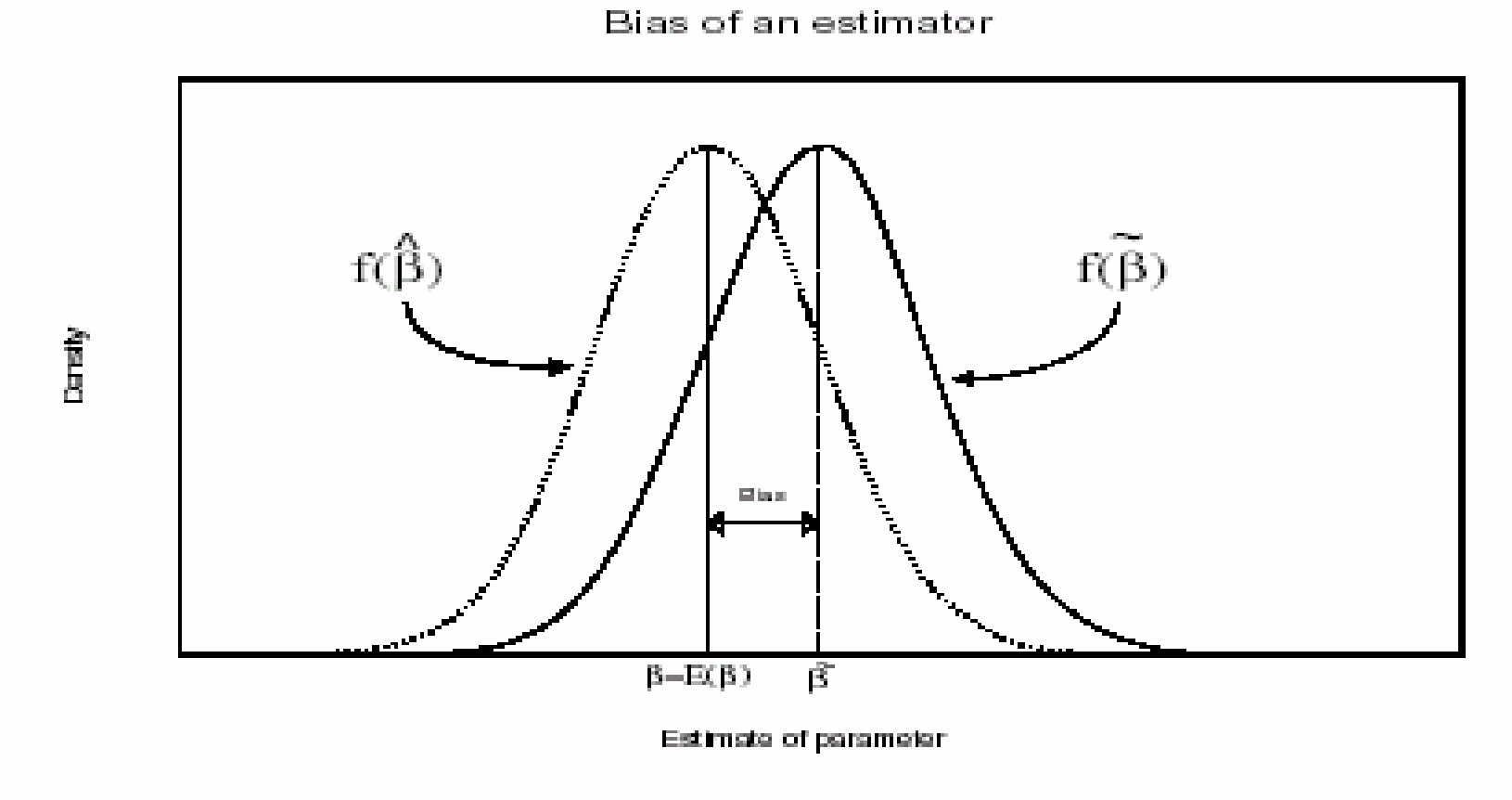
- Each additional year in education adds £1,450 to earnings
- Each additional year of experience adds £2,630 to earnings

1. Consequences of non-linearities

- Depending on how severe the non-linearity is, a, and b will be misleading:
  - estimates may be "biased"
  - i.e. they will not reflect the "true" values of α, β



$\beta^{\sim}$ is a biased estimator of β



2. Testing for non-linearities:

(a) visual inspection of plots

- scatter plots of two variables:
  - if you only have two or three variables then looking at scatter plots of these variables can help identify non-linear relationships in the data
  - but when there are more than 3 variables, non-linearities can be very complex and difficult to identify visually:
    - The effect of the higher dimensions cloaks the non-linear nature of the relationship between y and x

- What can appear to be random variation of data points around a linear line of best fit in a 2-D plot, can turn out to have a systematic cause when a third variable is included and a 3-D scatter plot is examined.
  - Same is true when comparing 3D with higher dimensions
  - e.g. Suppose that there is a quadratic relationship between x,y and z. But that this is only visible in the data if one controls for the influence of a fourth variable, w. But one does not know this, so looking at x, y and z, they appear to have a linear relationship.

2. Testing for non-linearities:
(b) t-statistics

- Sometimes variables that we would expect (from intuition or theory) to have a strong effect on the dependent variable turn out to have low t-values.
  - If so, then one might suspect non-linearities.
  - Try transforming the variable (e.g. take logs) and re-examine the t-values
    - e.g. HOUSING DEMAND $= a + b$ AGE OF BORROWER
      - surprisingly, age of borrower may not be that significant
      - but this might be because of a non-linearity: housing demand rises with age until mid-life, and starts to decrease as children leave home. Try Age^2 instead and check t-value.

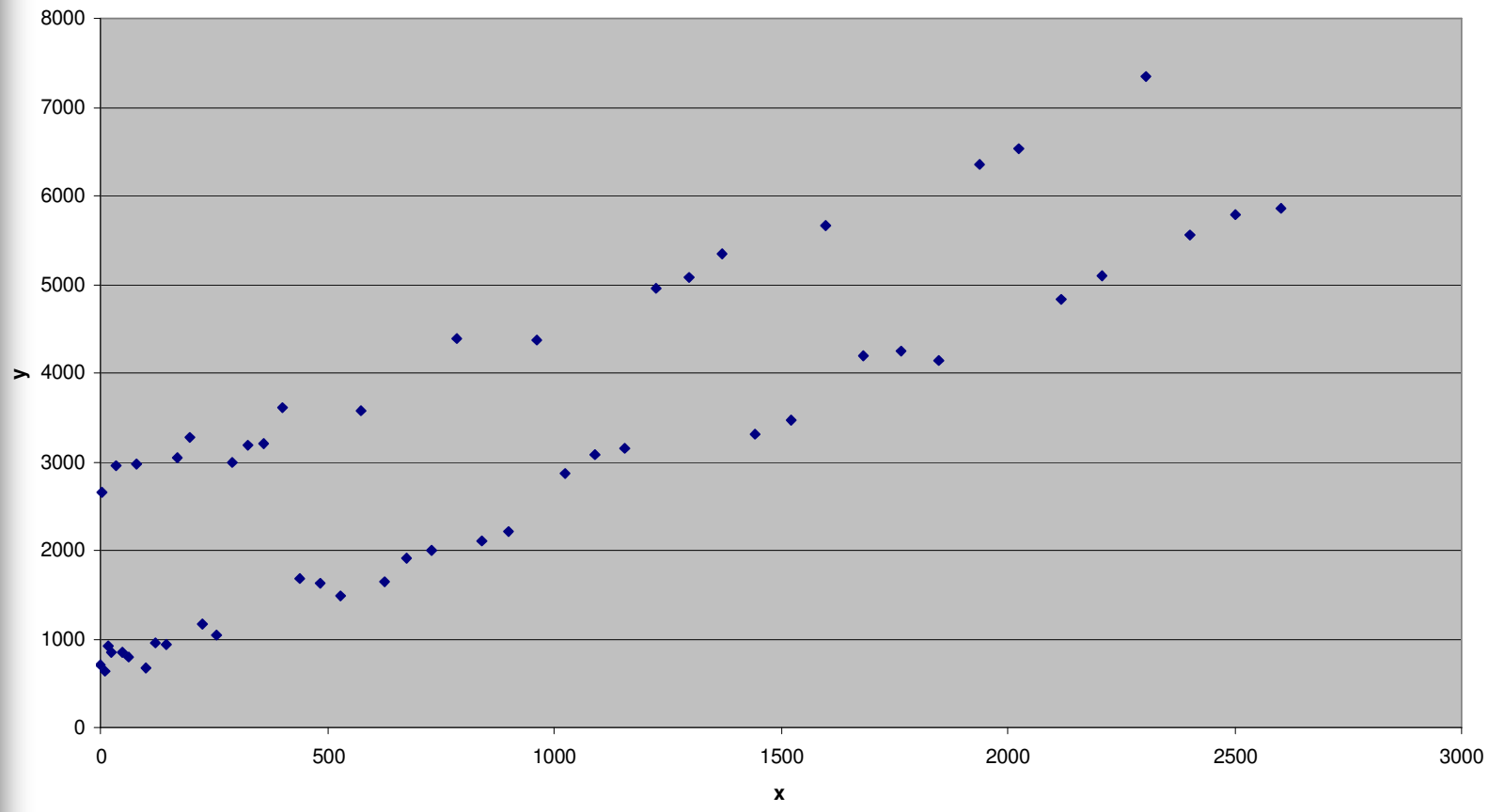
- There may be non-linearities caused by *interactions* between variables:
  - try interacting explanatory variables and examining t-values
    - e.g. HOUSE PRICE = *a* + *b* SIZE OF WINDOW + *c* VIEW
    - But size of window may only add value to a house if there is a nice view, and having a nice view may only add value if there are windows.
    - Try including and interactive term as well/instead:
      - HOUSE PRICE = *a* +…+ *d* SIZE OF WINDOW * VIEW
    - In SPSS you would do this by creating a new variable using the COMPUTE command:
      - COMPUTE SIZE_VEW = SIZE OF WINDOW * VIEW

      and then including the new variable in the regression.

2. Testing for non-linearities:
(c) shifts & structural break tests

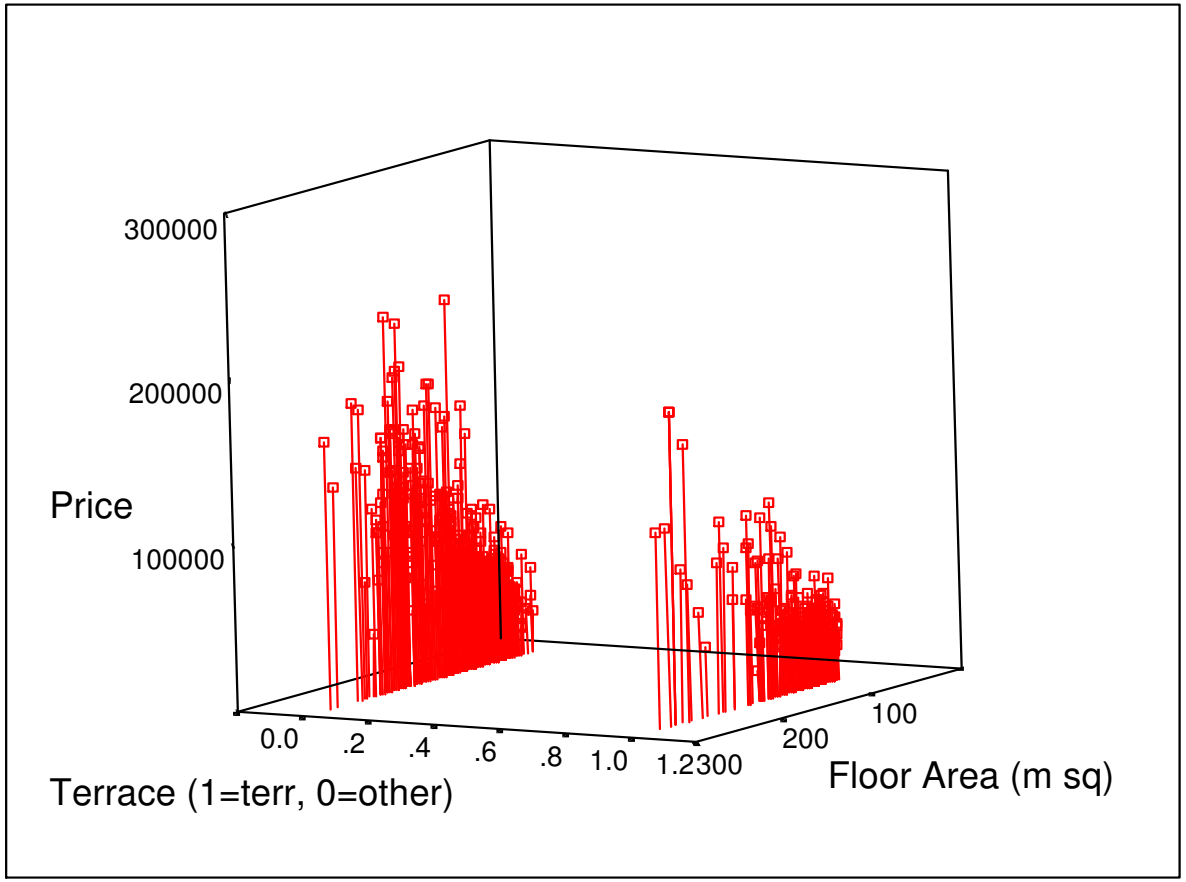
- Sometimes certain observations display consistently higher y values.
- If this difference can be modelled as a parallel shift of the regression line, then we can incorporate it into our model simply by including an appropriate dummy variable
  - e.g. male = 1 or 0;

Apparent Intercept Shift in data:



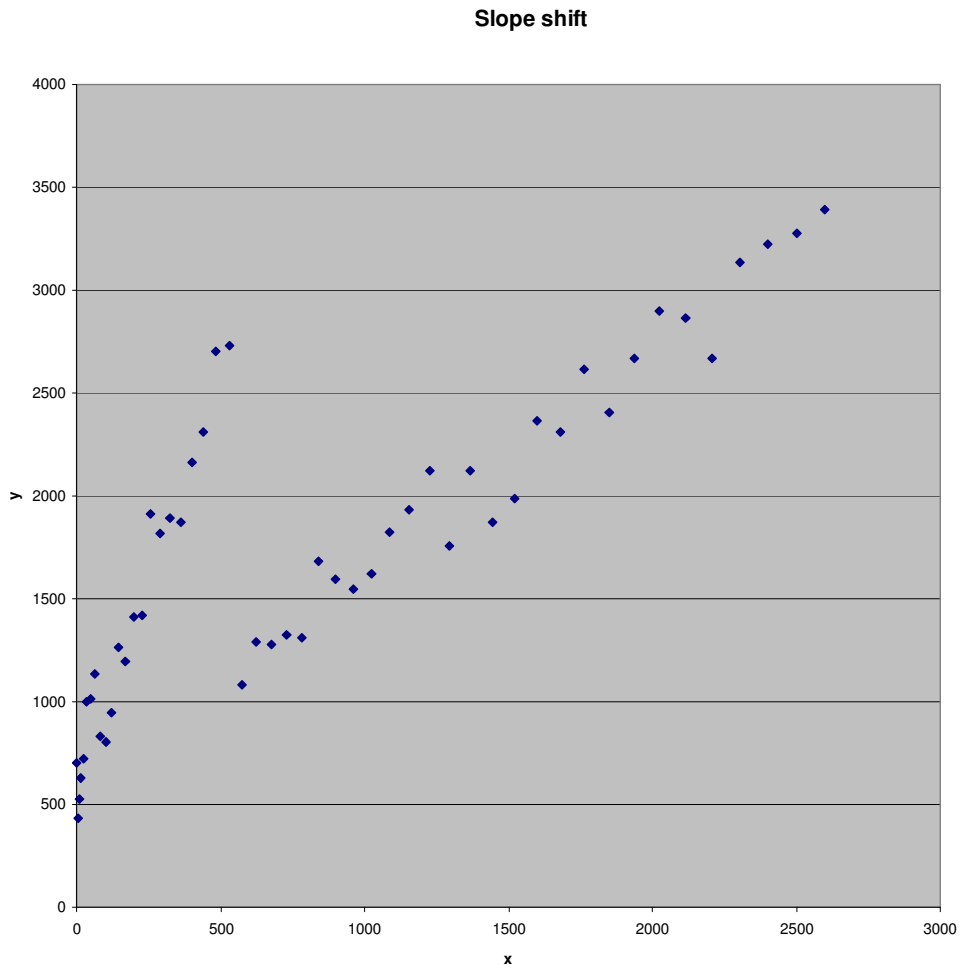
Data shifts in 3-Dimensions:

(NB: the shift is the slightly lower prices for terrace = 1)



- However, sometimes there is an apparent shift in the *slope* not just/instead of the intercept.
- Being able to observe this visually is difficult if you have lots of variables since the visual symptoms will only reveal themselves if the data has been ordered appropriately.

Apparent slope shift:



Solutions:

(a) Transforming Variables

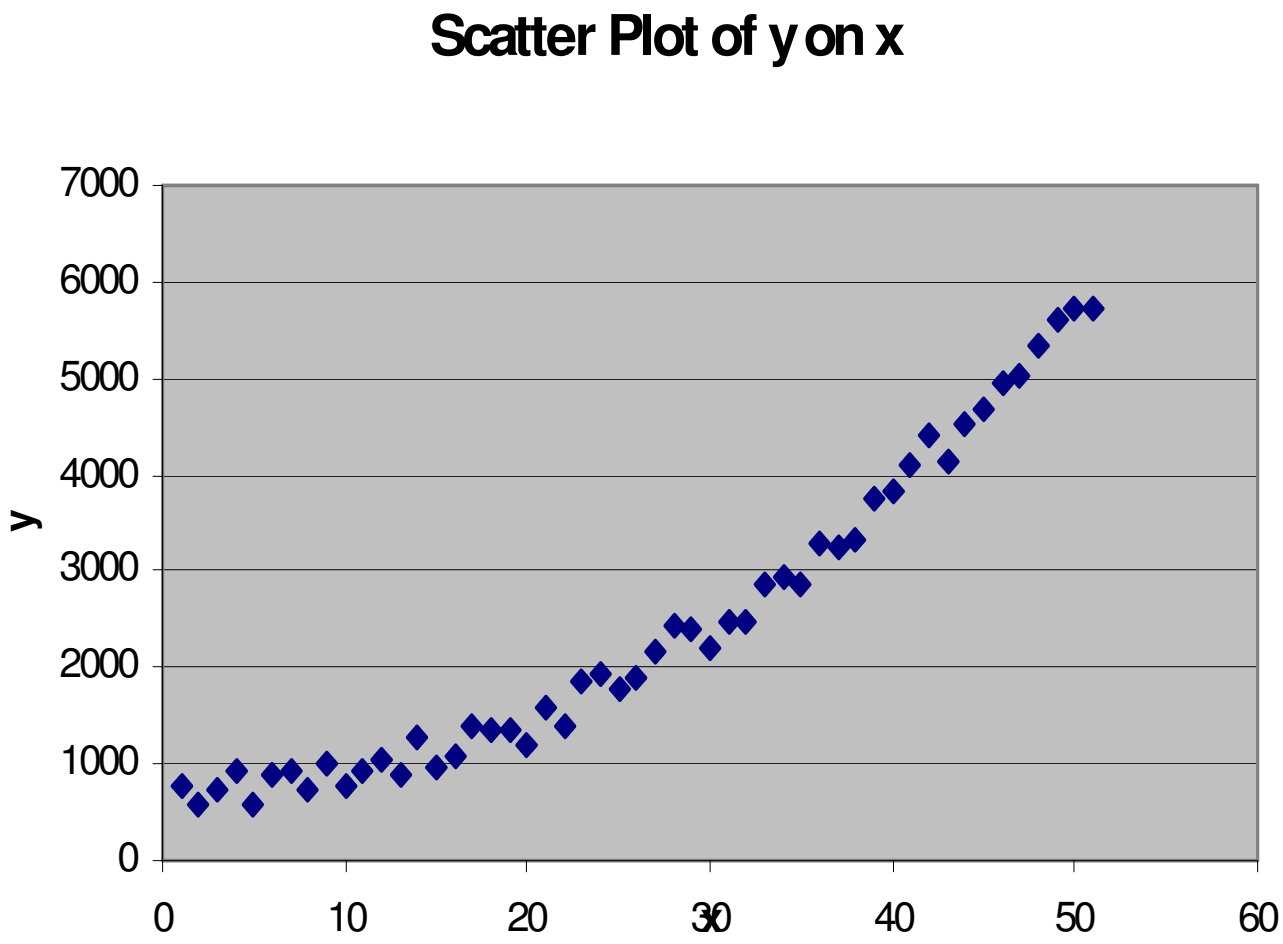
- Note that "Linear" regression analysis does not preclude analysis of non-linear relationships (a common misconception).
  - It merely precludes estimation of certain types of non-linear relationships
    - I.e. those that are non-linear in parameters:
    - $y = a^x + a^z + b^{xz}$

- However, so long as the non-linearity can fit within the basic structure of
  - $y = a + bx$
  - I.e. it is linear in parameters

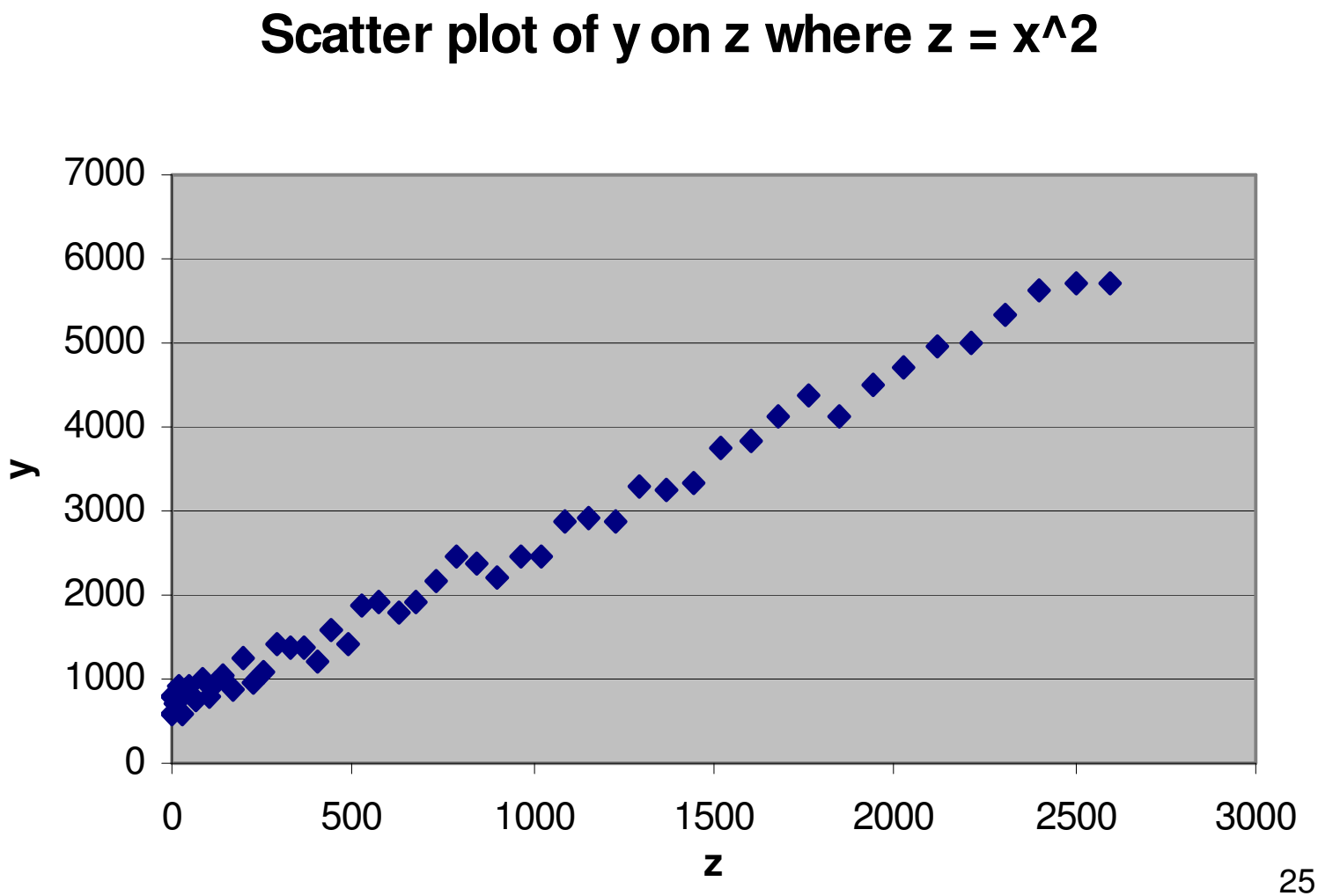
  – then we can make suitable transformations of the variables and estimate by OLS:

- e.g. 1 $y = a + b\,x^2$
  - we can simply create a new variable, $\mathbf{z = x^2}$ and run a regression of $y = a + b\,\boldsymbol{z}$
  - including the square of x is appropriate if the scatter plot of y on x is “n” shaped or “u” shaped
- e.g. 2 $y = b + bx^3$
  - we can create a new variable, $\mathbf{z = x^3}$ and run a regression of $y = a + b\,\boldsymbol{z}$
  - including the square of x is appropriate if the scatter plot of y on x is “s” shaped or has a back-to-front “s” shape.

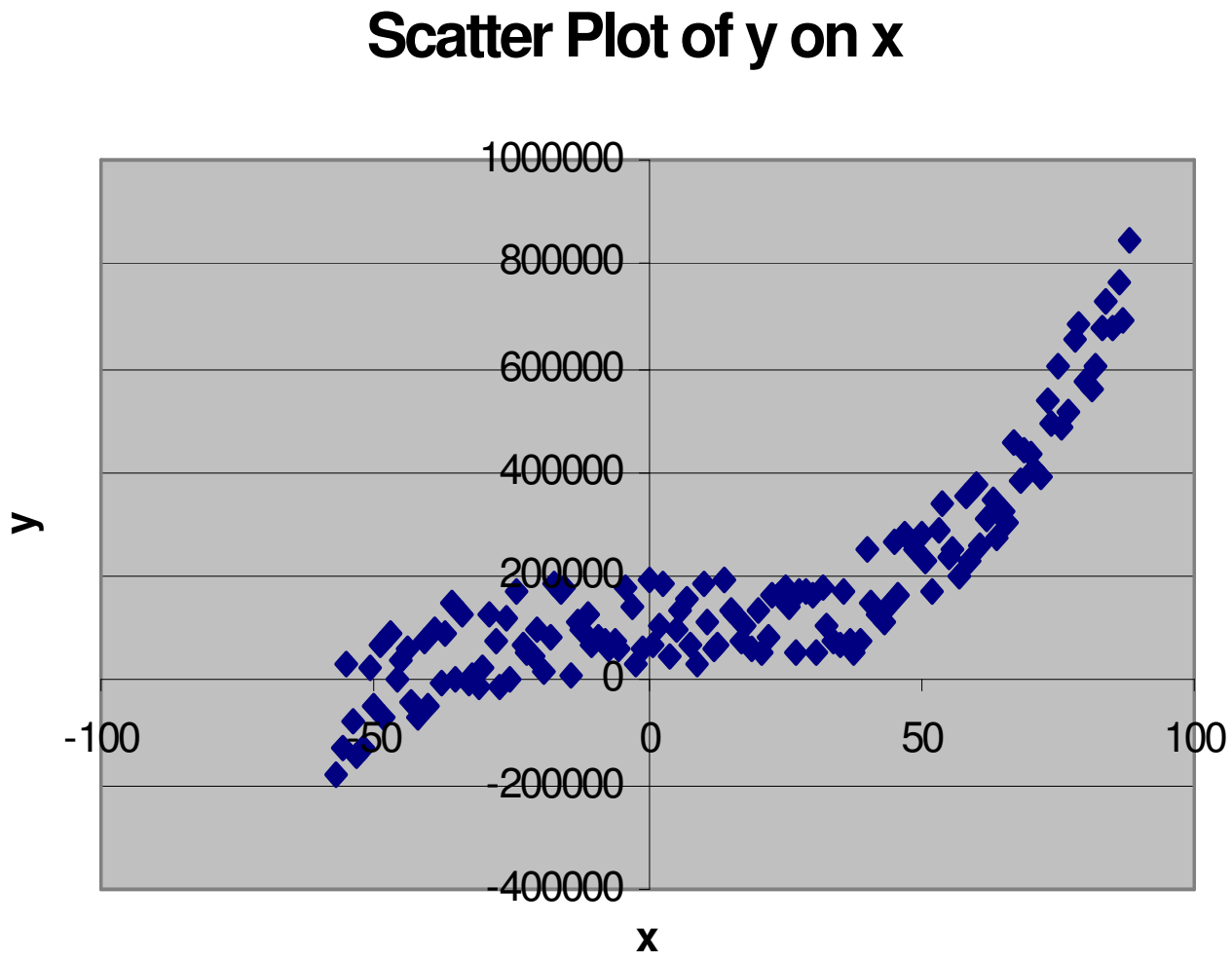
E.g.1 Scatter plot suggests a quadratic relationship



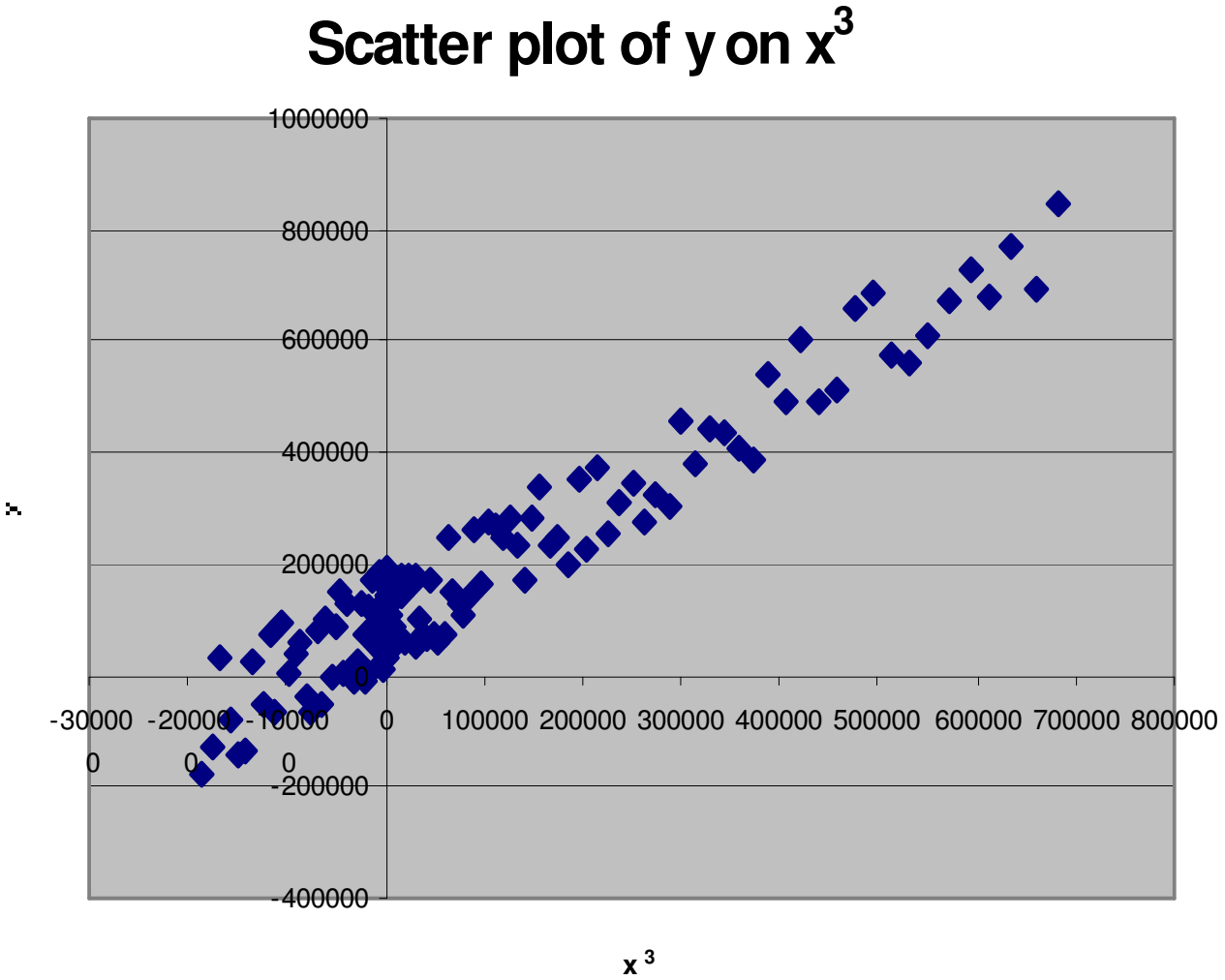
Regressing *y* on the square of *x* should give a better fit



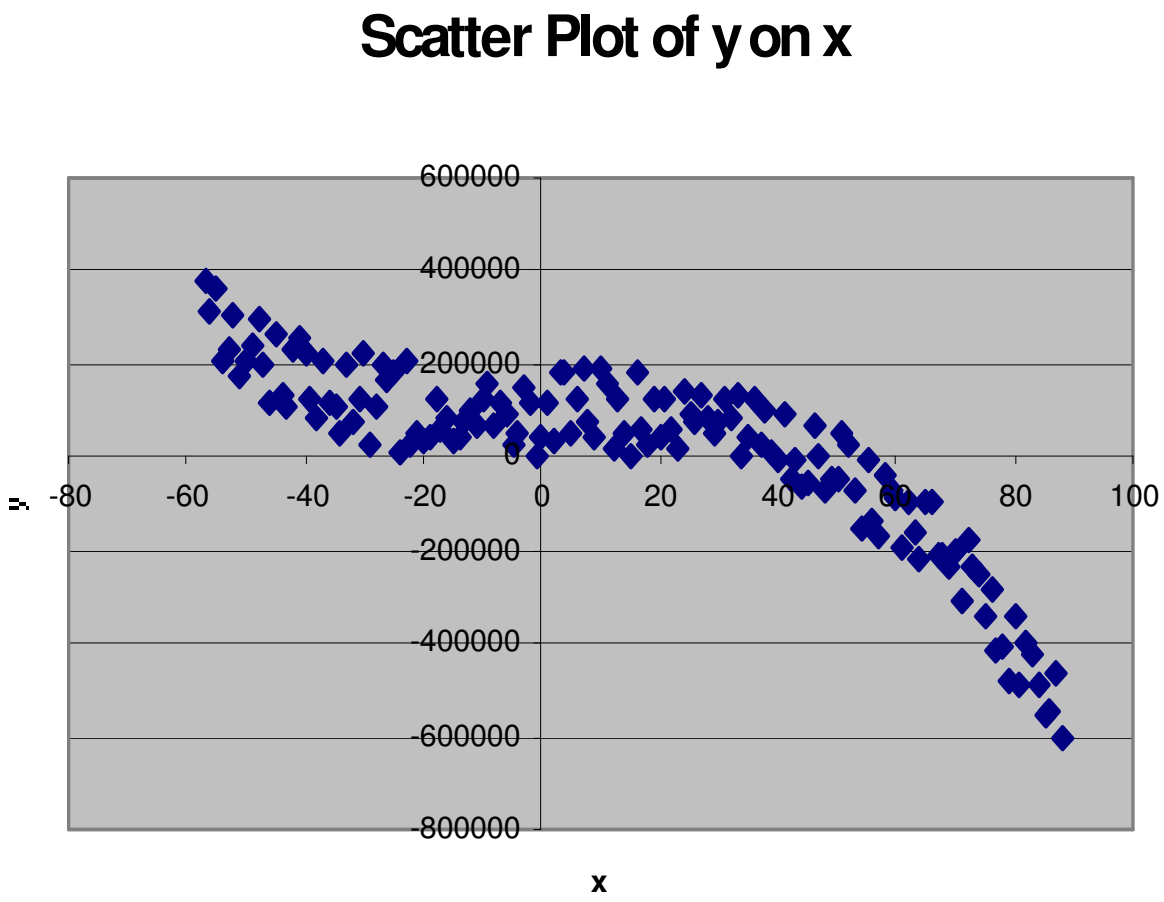
E.g. 2 Scatter plot suggests a cubic relationship



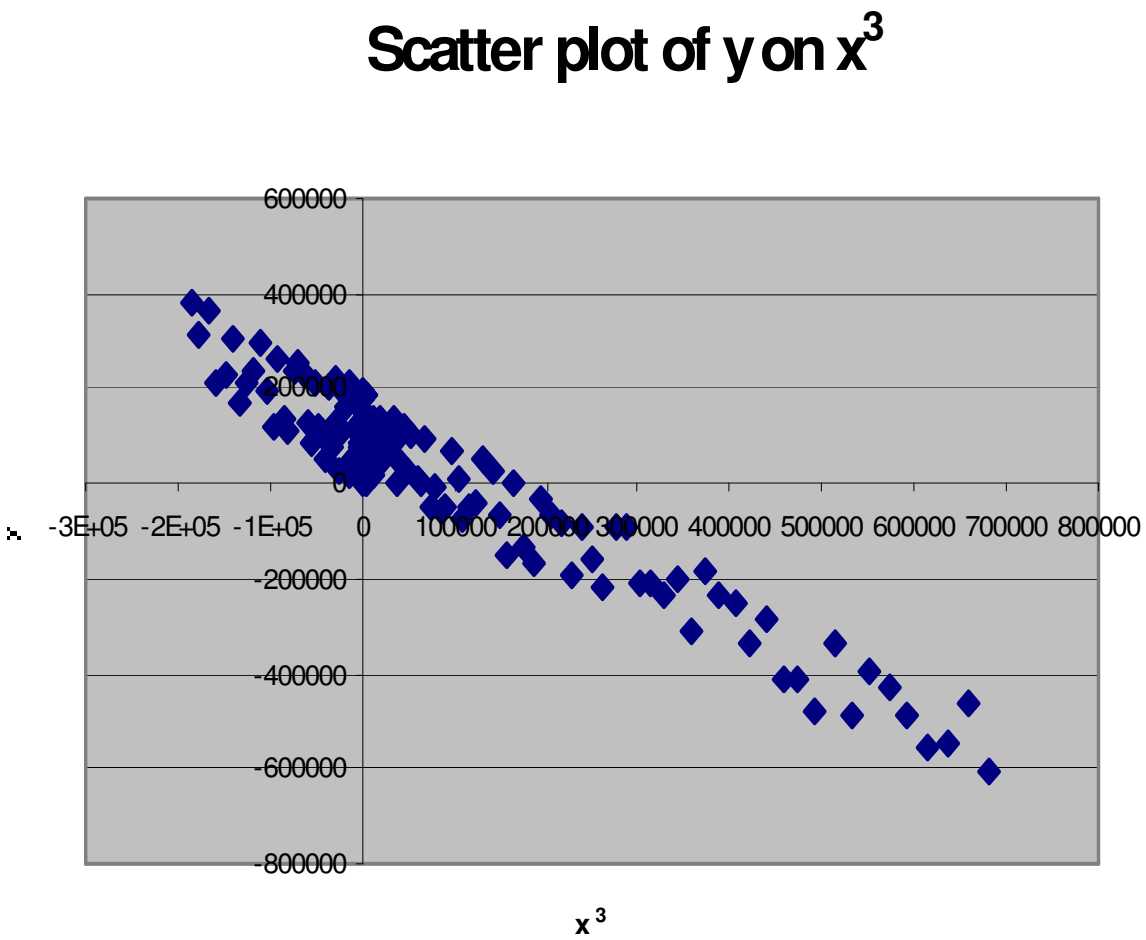
Regressing *y* on the cube of *x* should give a better fit:



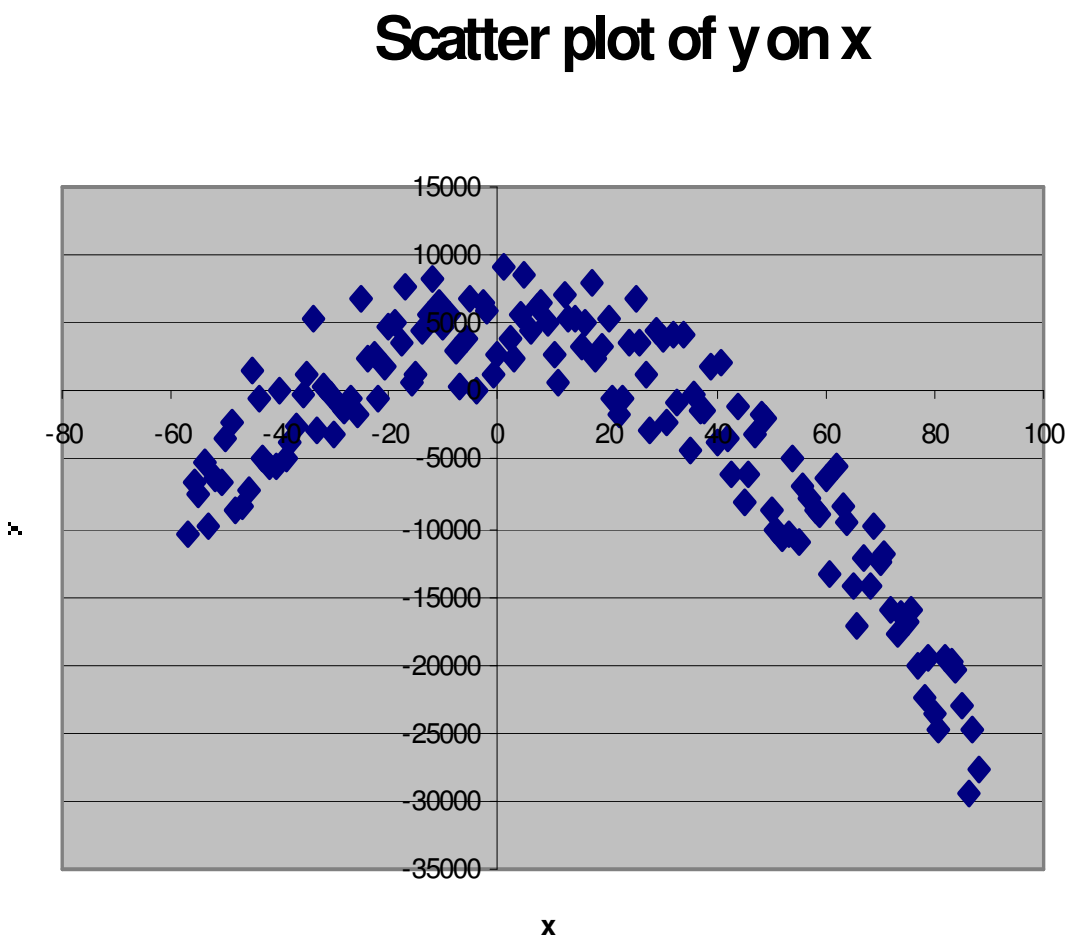
E.g. 3 Scatter Plot suggests a cubic relationship



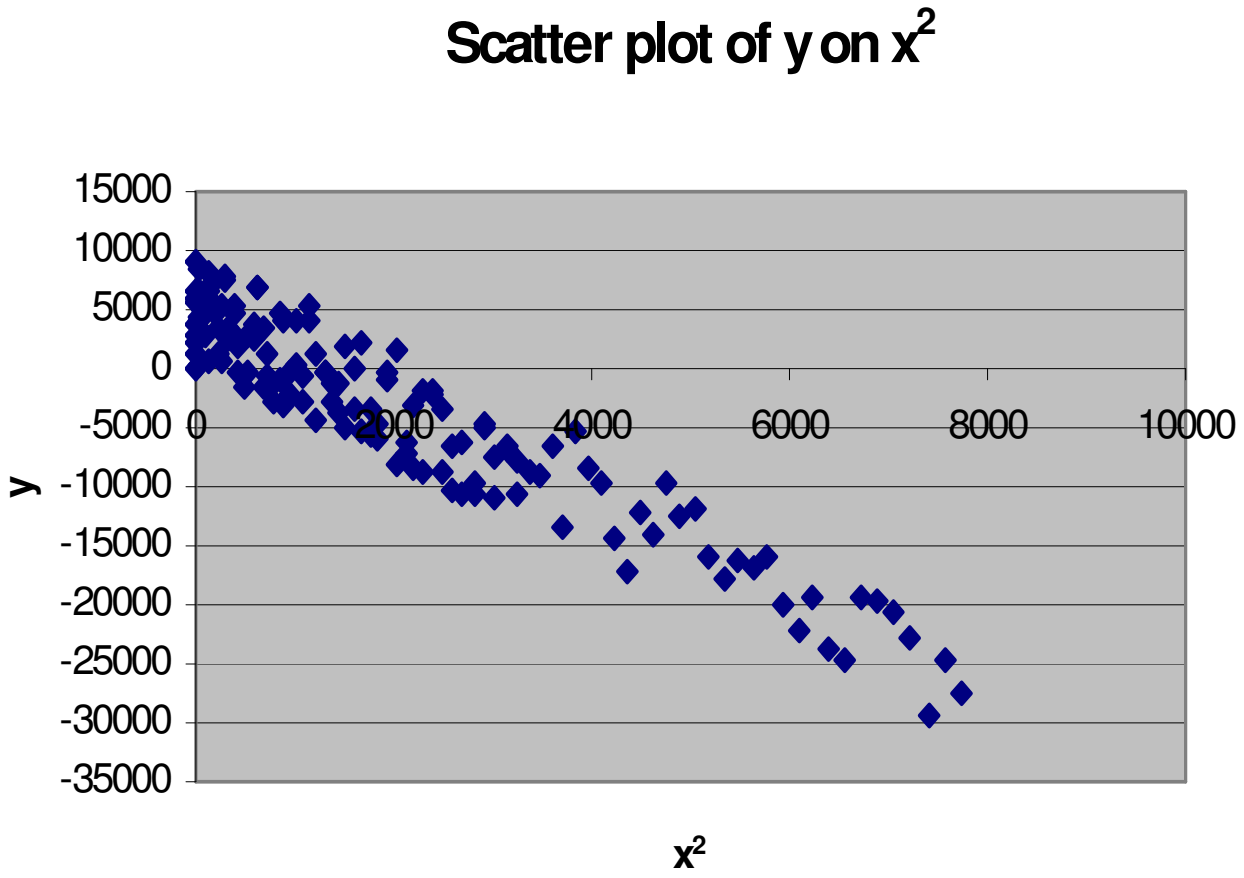
Cubing x should give a better fit



E.g. 4 Scatter plot suggests a quadratic relationship



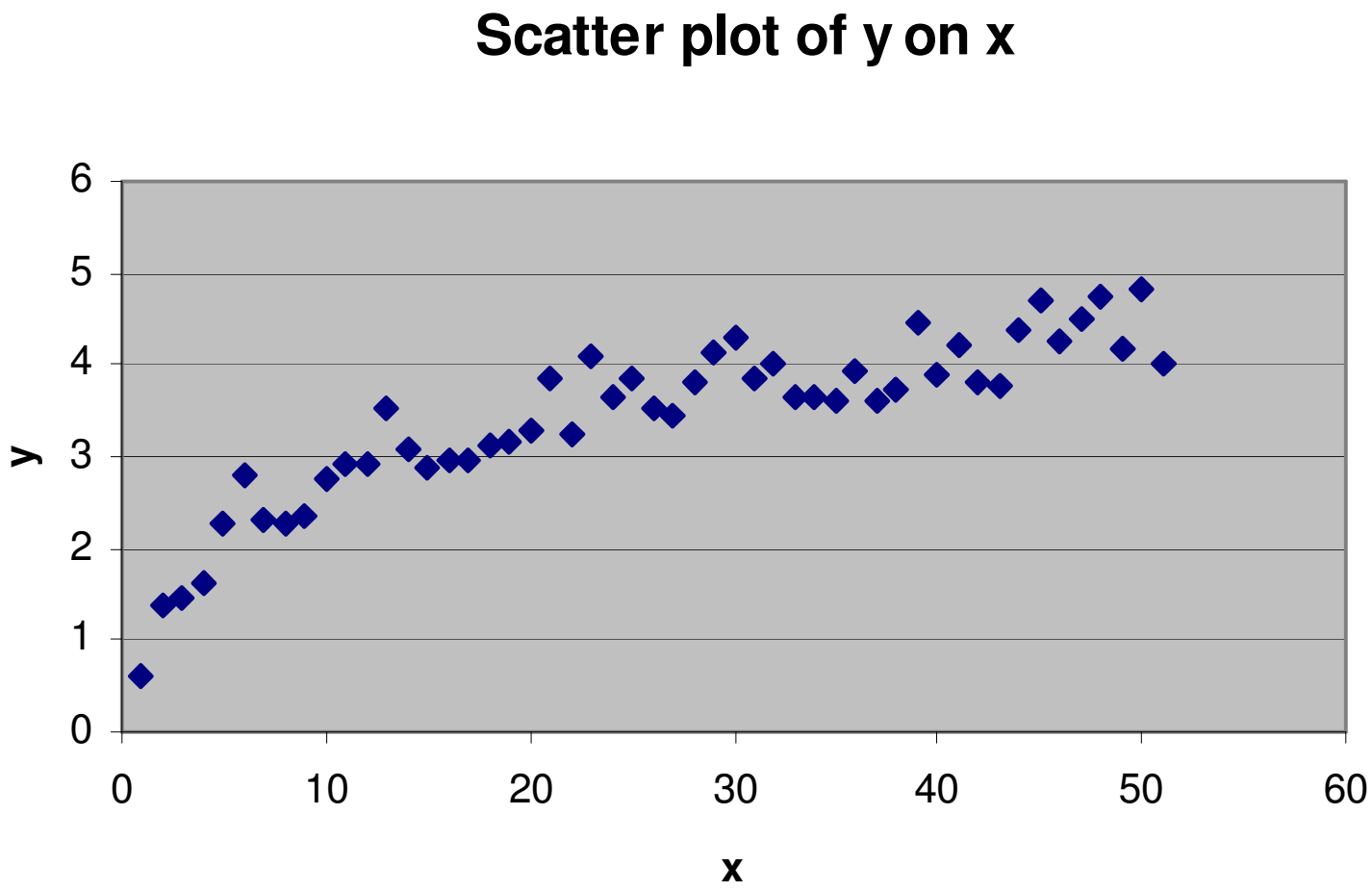
Squaring *x* should give a better fit



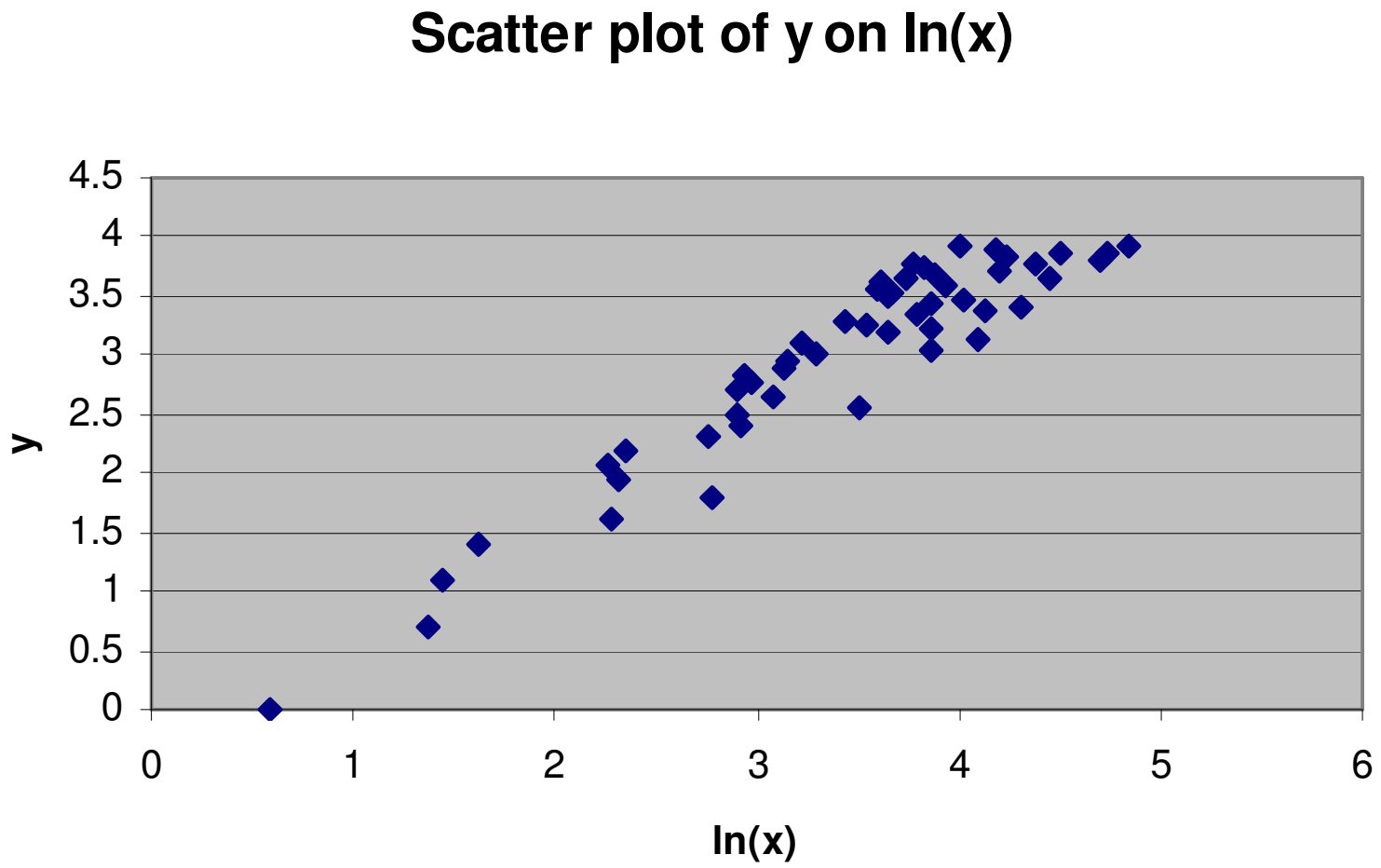
Log-log and log-linear models

- One of the most common transformations of either the dependent variable and/or the the explanatory variables is to take logs.
  - It is appropriate to transform *x* if the scatter plot of *y* on *x* has either an “r” shape, or an “L” shape.

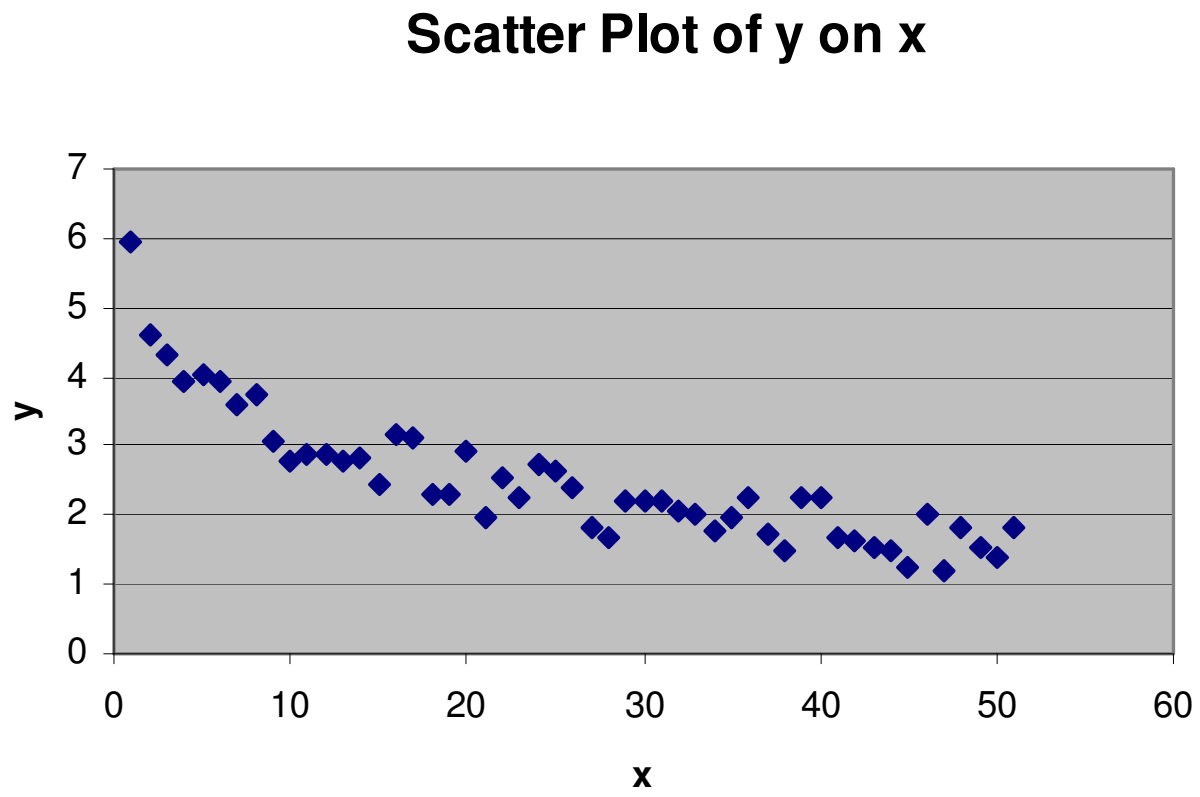
E.g 5 scatter plot suggests a logarithmic relationship



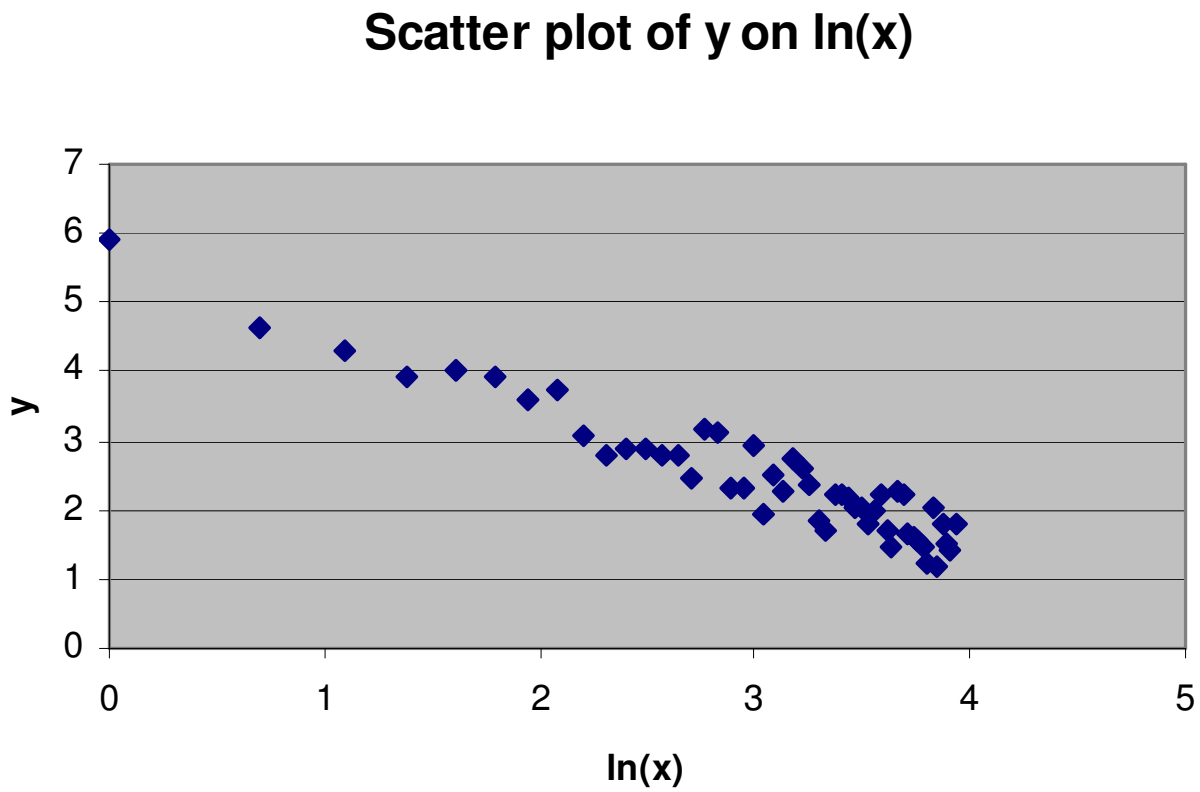
Taking the log of *x* should result in a better fit



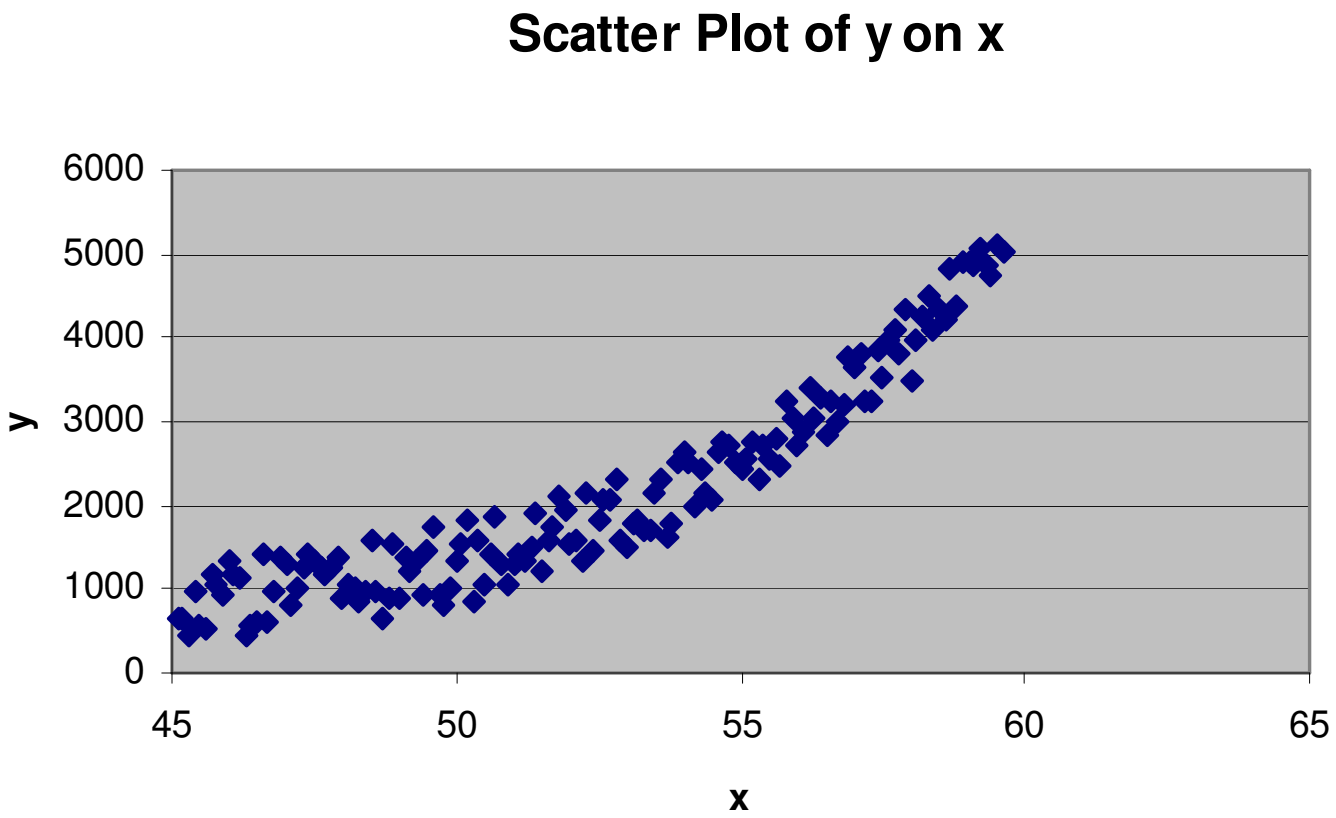
E.g.6 scatter plot suggests a logarithmic relationship



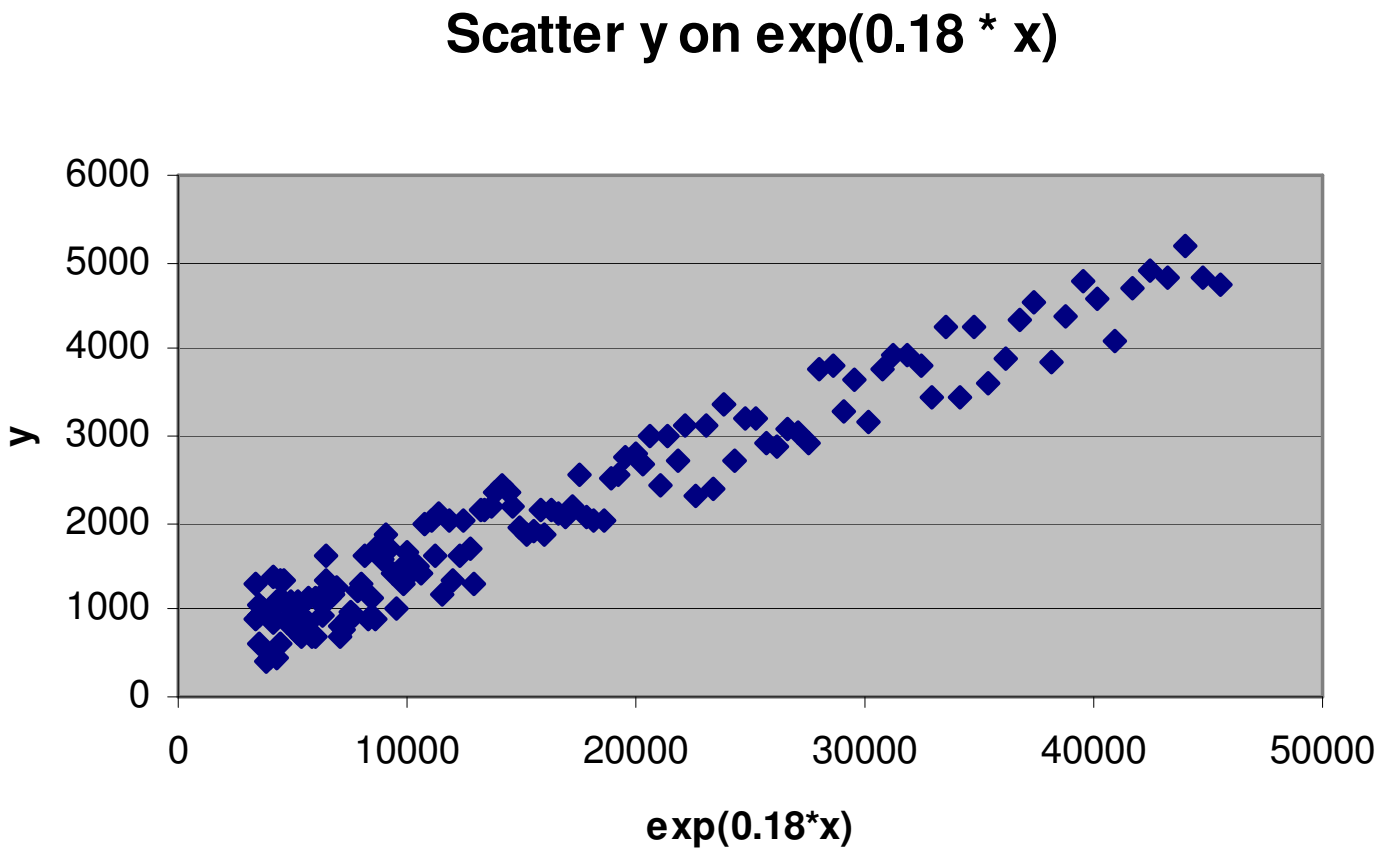
Taking the log of *x* should result in a better fit



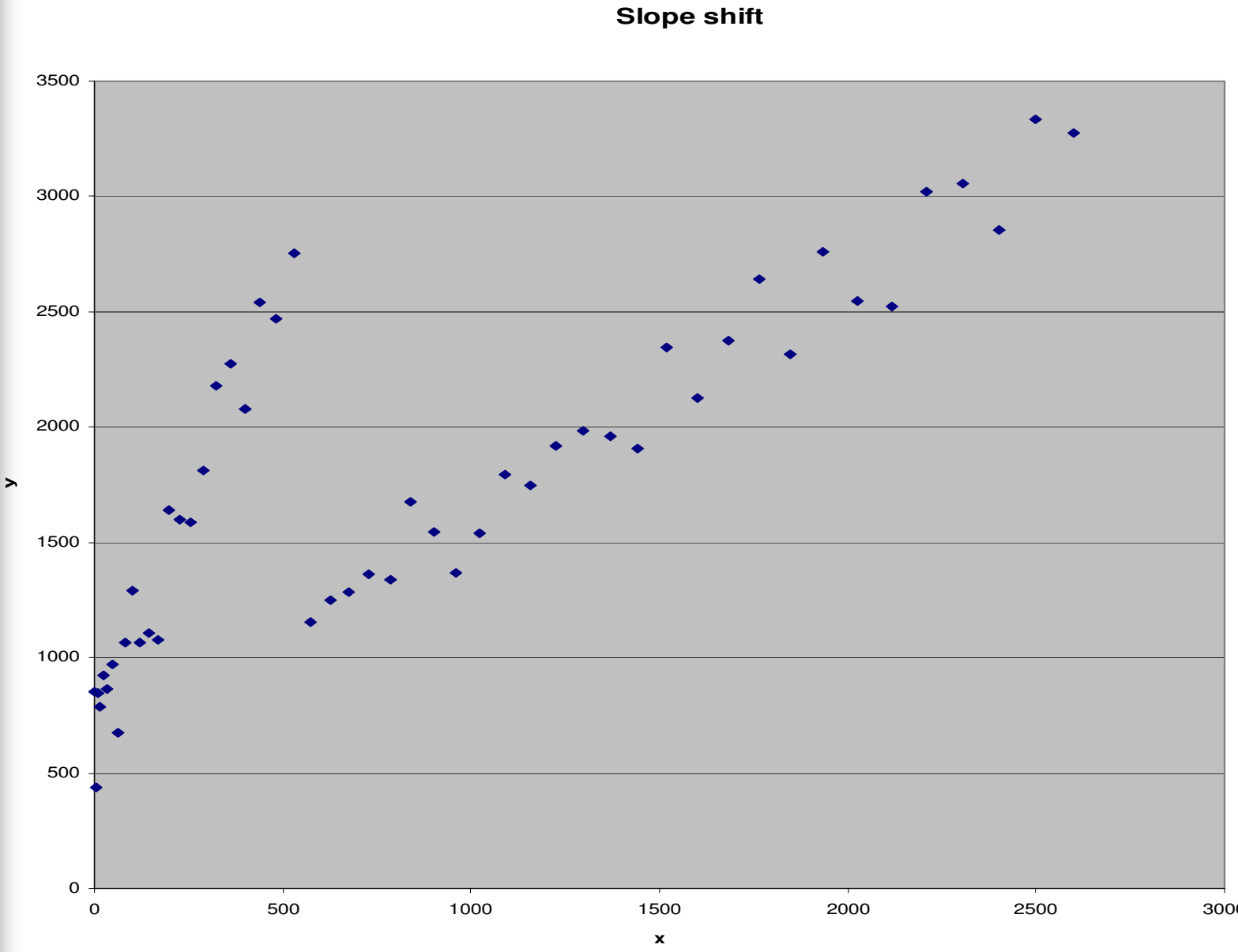
E.g. 7 scatter plot suggests an exponential relationship



Taking the exponent of *x* should result in a better fit



Solutions:b) Split the sample



- Quite a drastic measure:
  - split the sample and estimate two OLS lines separately
  - in practice its not easy to decide where exactly to split the sample
  - we can do an F-test to help us test whether there really is a structural break: “Chow Tests”
  - but even if the F-test shows that there is a break, it can often be remedied by squaring the offending variable, or using slope dummies...

Solutions:
(c) Dummy variables

- A dummy variable is one that takes the values 0 or 1:
  - e.g. 1 if male , 0 if female
- If we include the dummy as a separate variable in the regression we call it an *Intercept Dummy*
- If we multiply it by one of the explanatory variables, then we call it a *Slope Dummy*

Intercept Dummies:

Original equation:

$y = a + bx$

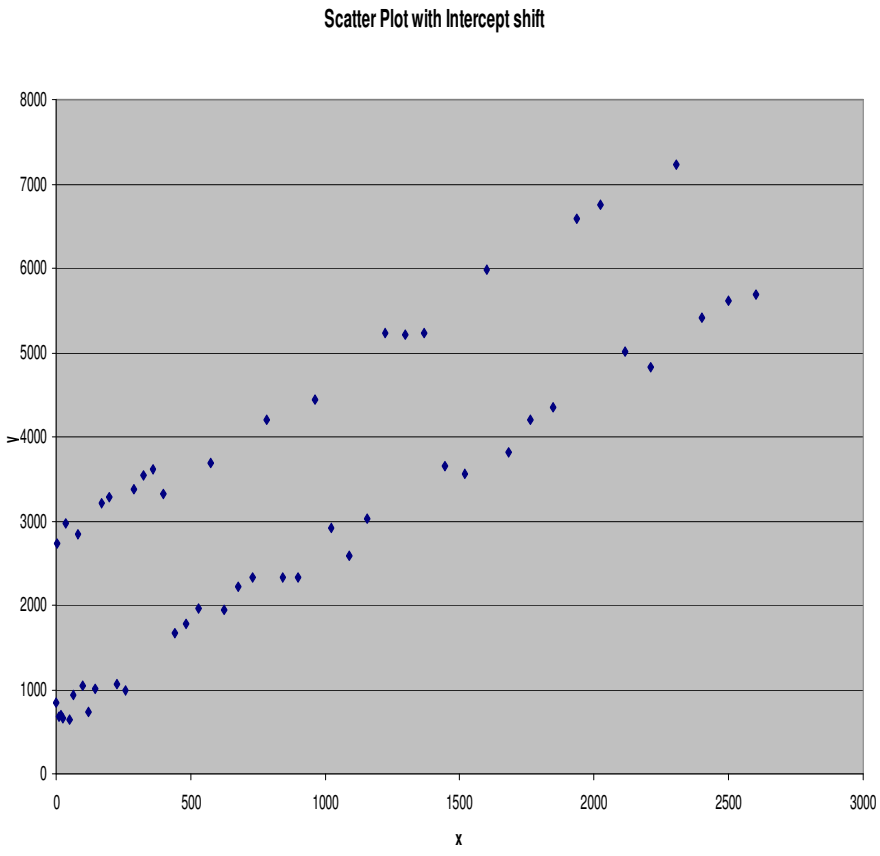
now add a dummy:

(eg.D= 0 if white,

D= 1 if non-white)

$y = a + bx + cD$

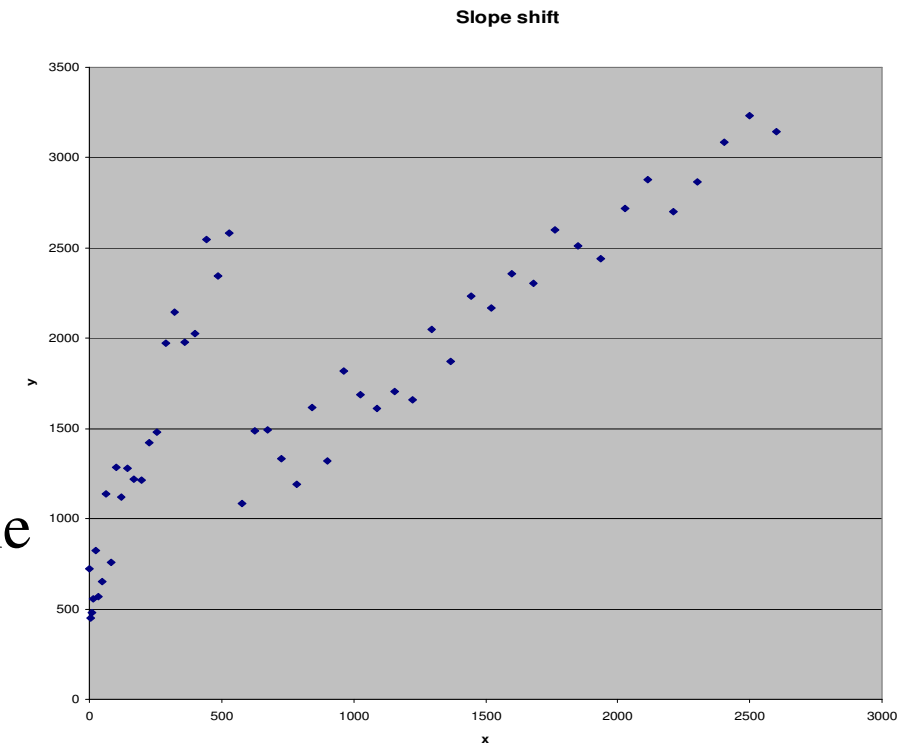
c measures how much higher (lower if c is negative) the dependent variable is for non-whites



Slope Dummies:

Suppose race has an effect on the slope of the regression line rather than the intercept.

You can account for this by simply multiplying the relevant explanatory variable by the race dummy:



$$y = a + bx + cD*x$$

c measures how much higher (lower if c is negative) the b slope parameter would be for non-whites

Solutions:
(d) Non-linear estimation

- When you can't satisfactorily deal with the non-linearity by simply transforming variables, you can fit a non-linear curve to the data
- These are usually based on some sort of grid search (I.e. trial and error) for the correct value of the non-linear parameter.
  - E.g. $y = a + b_1 e^{b_2 x + b_3 z}$
    - cannot be transformed to linearity in a way that would allow us to derive estimates for b_2 and b_3

- SPSS does allow non-linear estimation
  - go to Analyse, Regression, non-linear
- But we shall not cover this topic in any more detail on this course since most types of non-linearity in data can be adequately dealt with using transformations of the variables.

Summary

- 1. Consequences of non-linearities
- 2. Testing for non-linearities
  - (a) visual inspection of plots
  - (b) t-statistics
  - (c) structural break tests
- 3. Solutions
  - (a) transform variables
  - (b) split the sample
  - (c) dummies
  - (d) use non-linear estimation techniques

Reading:

- Appendix A.4 “Some Special Functions and their Properties”, Wooldridge, Introductory Econometrics, Third Edition.
- Kennedy (1998) “A Guide to Econometrics”, Chapters 3, 5 and **6**