Unit 5: Regression

Marius Ionescu

09/22/2011

SD line

Fact

- *The scatter plot generally clusters about the SD line, but does not, in general, give the line of linear correlation.*
- *It goes through the ends of the football, so gives the direction of the football.*
- *Not used for predictions, only for location purposes on a scatterplot.*
- *Goes through* (Avg_x, Avg_y) *with slope* $\pm SD_y/SD_x$ *depending on* $r > 0$ *or* $r < 0$.
- *Hence, the line is*

$$y = \pm \frac{SD_y}{SD_x}(x - Avg_x) + Avg_y.$$

SD line

Fact

- *Has nothing to do with r (aside from $r < 0$ or $r > 0$), so we can have same SD-line for sets of data with identical averages, SDs, regardless of respective r values.*
- *Clearly, this is not what we want.*

Fact

Regression=*Attempt to quantify how one variable depends on another given a certain degree of correlation.*

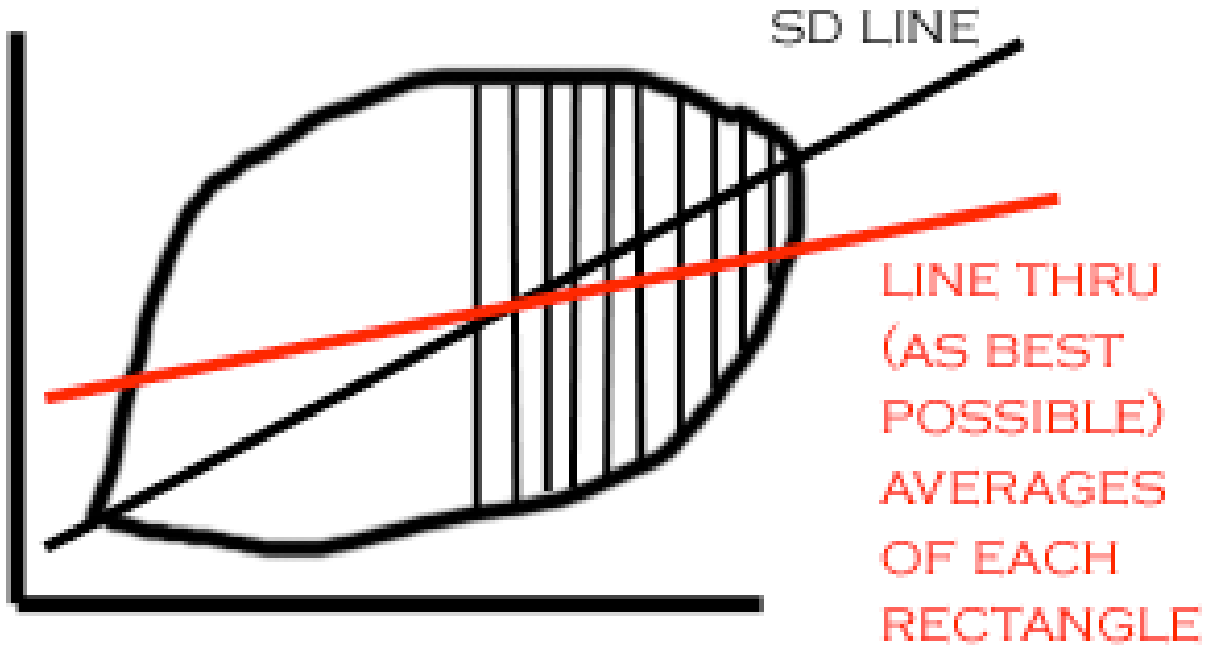
Example

Example

Suppose that we measure the relationship between height and weight. Let X =height, Y =weight. We have $Avg_x = 70$ in., $Avg_y = 162$ lbs., $SD_x = 3$ in., $SD_y = 30$ lbs., with $r = .47$.

- Q: So if we know someone is 73 inches tall, what we predict for their weight?
- A: **Average** weight of all 73 inch tall people in the study.
- How do we get this?

The regression line (best fit line)



Fact

The line in red is called the regression line for predicting y given x.

The regression line

Fact

- *If the data is nice (normally distributed and linear correlated), then the average of y on each rectangle will be on a line.*
- *Graphically the regression line goes through the Avg point* ***and*** *the sides of the football.*

The regression line

Definition

- The regression line is given by the formula

$$y - Avg_y = r\frac{SD_y}{SD_x}(x - Avg_x).$$

- Note that the variable being predicted is on the left.
- By construction it goes through the point (Avg_x, Avg_y).

The regression line in standard units

Fact

- *We can rewrite the regression line as*

$$\frac{y - Avg_y}{SD_y} = r\frac{x - Avg_x}{SD_x}.$$

- *Thus, in standard units,*

$$Y = rX.$$

Example

Example (Weight vs Height, cont'd)

What is the average weight of those 73" tall?

Another example

Example

Suppose that Colgate students who take the LSAT have $Avg_{LSAT} = 600$, $SD_{LSAT} = 50$, $Avg_{GPA} = 3.3$, $SD_{GPA} = 0.3$ and the correlation coefficient $r = 0.6$. What is the regression prediction for the LSAT score of the students with GPA of

- 3.3?
- 3.6?

The regression fallacy

Fact

- *A predicted value of y will be close to the average of y (in terms of SDs) than the x value is to the average of x (in terms of SDs).*
- *In virtually all test-retest situations, the bottom group's score will, **on average**, improve and the top group's scores will, **on average**, decline.*

Example

Example (Are sophomores dumber?)

Suppose that the average GPA for the first and second year students at Colgate it 2.8, SD is 0.7, and the correlation coefficient is $r = 0.8$. What is the most likely outcome for the best student (4.0 GPA as a first year student) in the second year?

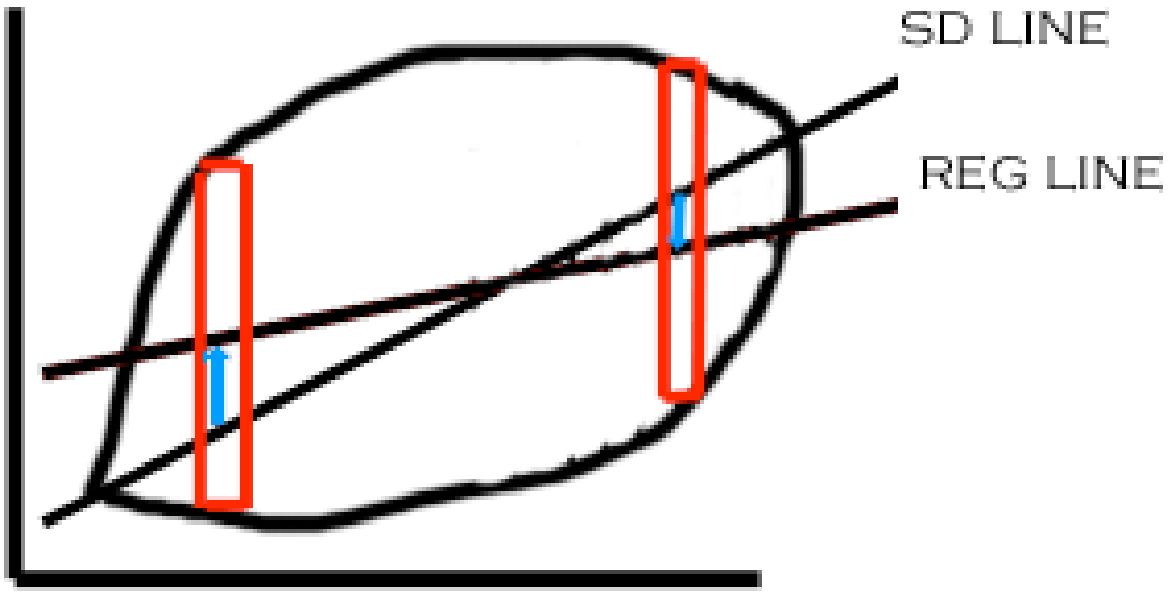
The regression fallacy

Fact

Remember:

- *Regression is an idealized description of a population.*
- *It doesn't describe individual cases.*
- *The previous example does not say that sophomores do worse than in the first year.*
- *It says that the best 1^{st} years will on average do worse in their 2^{nd} year, even though the overall average and SD stay the same.*

The regression fallacy



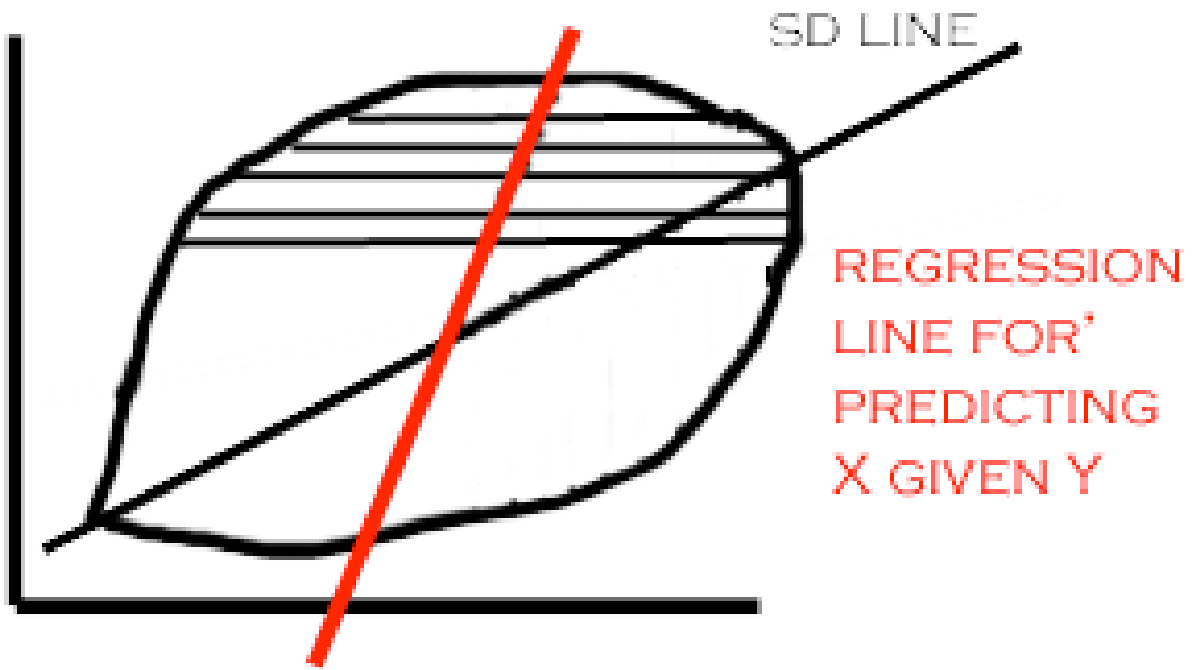
Fact

Blue arrows show increase in scores for those below average and decrease in scores for those above average.

There are TWO regression lines

Fact

One for predicting y given x; one for predicting x given y.



Example

Example

Suppose that the average IQ is 100 and SD is 15 for both women and men and husband/wife couples have a correlation coefficient of 0.5.

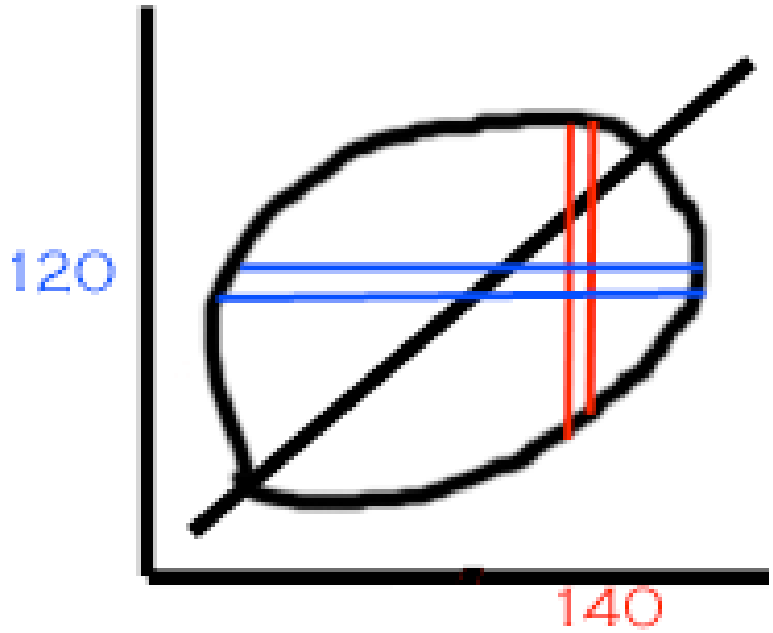
- What is the predicted average IQ of women whose husbands have IQ of 140?
- What do you predict for the average IQ of their husbands?

Example (cont'd)

Fact

- *So smart husbands have dumber wives, who in turn have dumber husbands.*
- *What gives?*
- *Again, different families of data under consideration:*
  1. *All wives whose husbands have IQs of 140;*
  2. *All husbands whose wives have IQs of 120:*

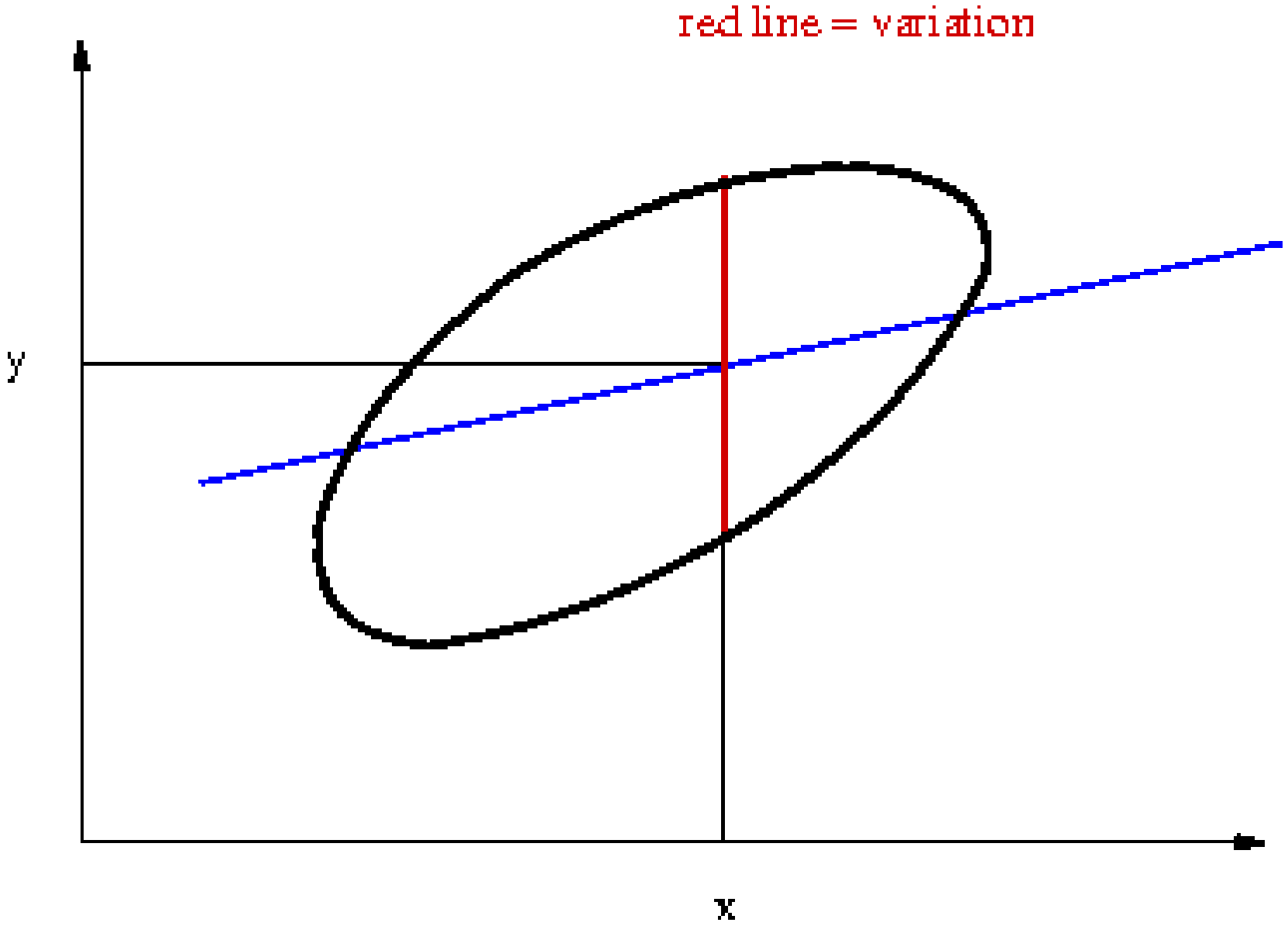
Example (cont'd)



Fact

- *Recall that if x and y are linearly correlated and $x = x_1$ then the value y predicted using the regression line is the* ***average*** *value*
- *What is the variation expected around this?*

Variation



The RMS error

Definition

The RMS error is

$$\sqrt{\text{Average of the squared errors.}}$$

Fact

- *The regression line is the line which minimizes the RMS error.*
- *For this reason the regression line is called also the* ***least square error*** *or* ***least squares*** *line.*

Properties

Fact

- *The RMS error is a vertical measure of the error.*
- *For ideal situation, 68% of data will fall within 1 RMS error of the regression line.*
- *95% of data will be within 2 RMS errors of the regression line.*
- *The RMS error behaves like the SD from the regression line.*

Formula for computing the RMS

Fact

- *We can compute the RMS using the formula*

$$RMS\ error = \sqrt{1 - r^2} \cdot SD_y.$$

- *This formula is valid for normally distributed linearly correlated data.*

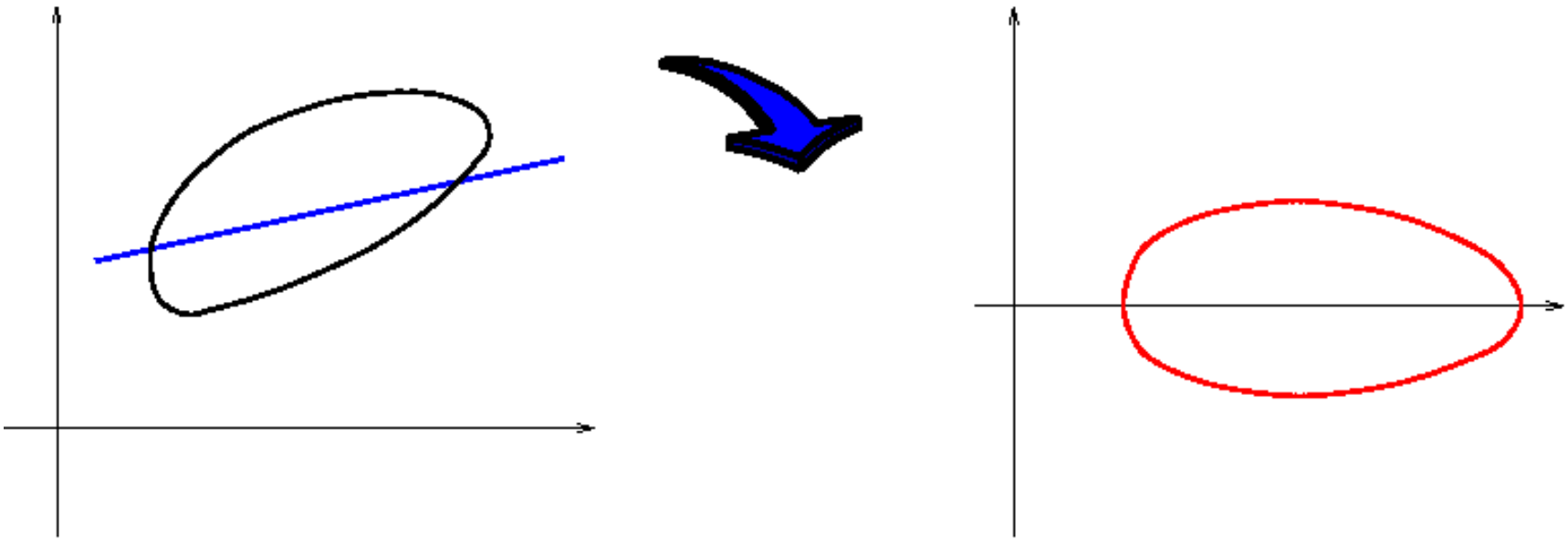
Example

Example

Suppose that for Colgate freshmen the average of the SAT verbal is 550, SD is 50, the average of their GPA is 2.8, SD is 0.7, and $r = 0.5$.

- What is the RMS error when predicting the SAT verbal?
- For students scoring 650 on the SAT verbal, what is the predicted GPA with error estimate? .

Residuals



Residuals

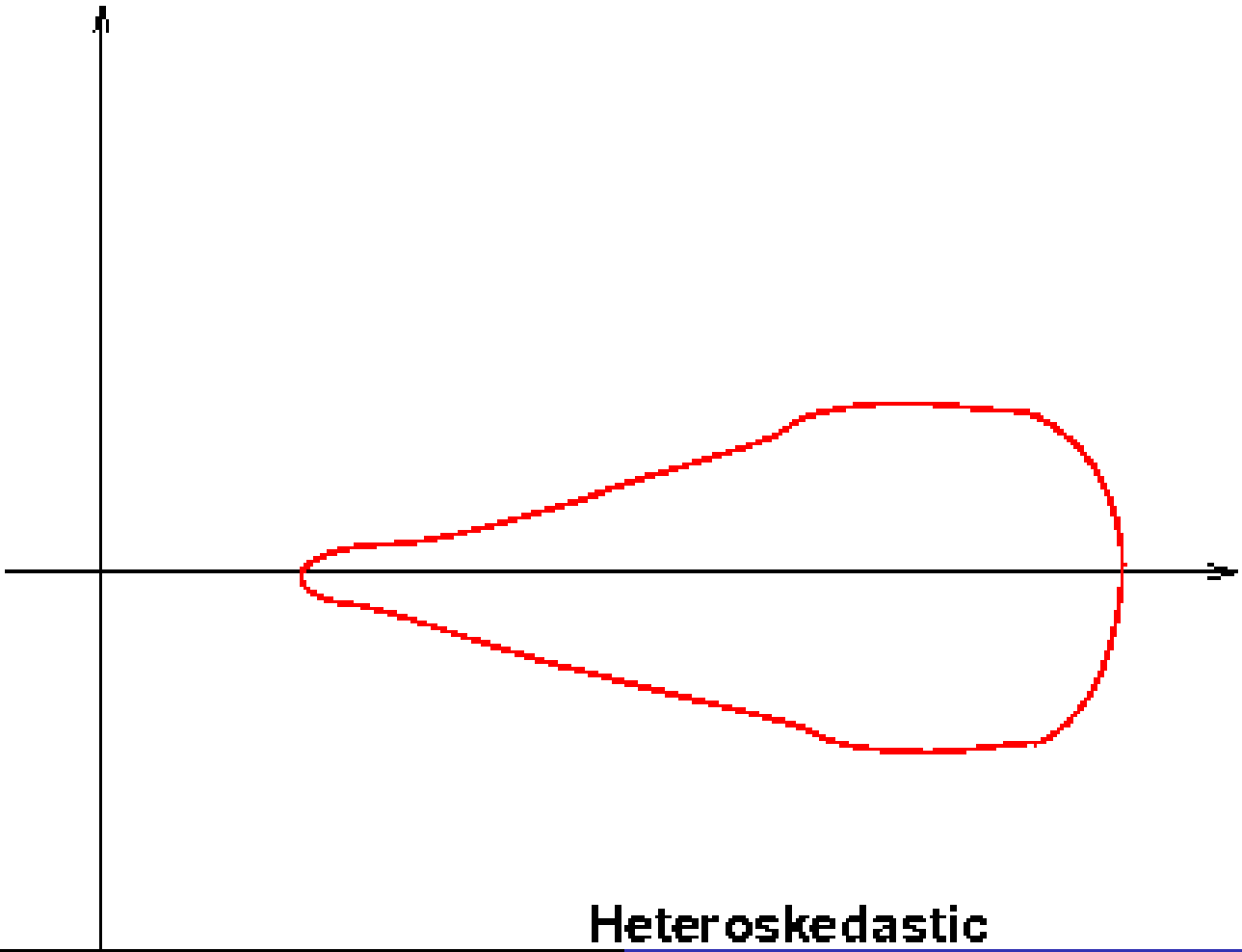
Fact

- *The residuals are the errors centered around* 0.
- *They should have* ***no*** *"linear trend" (that is, the slope should be* 0.

Heteroskedastic data

Definition

We say that data is heteroskedastic if the residuals have different distributions

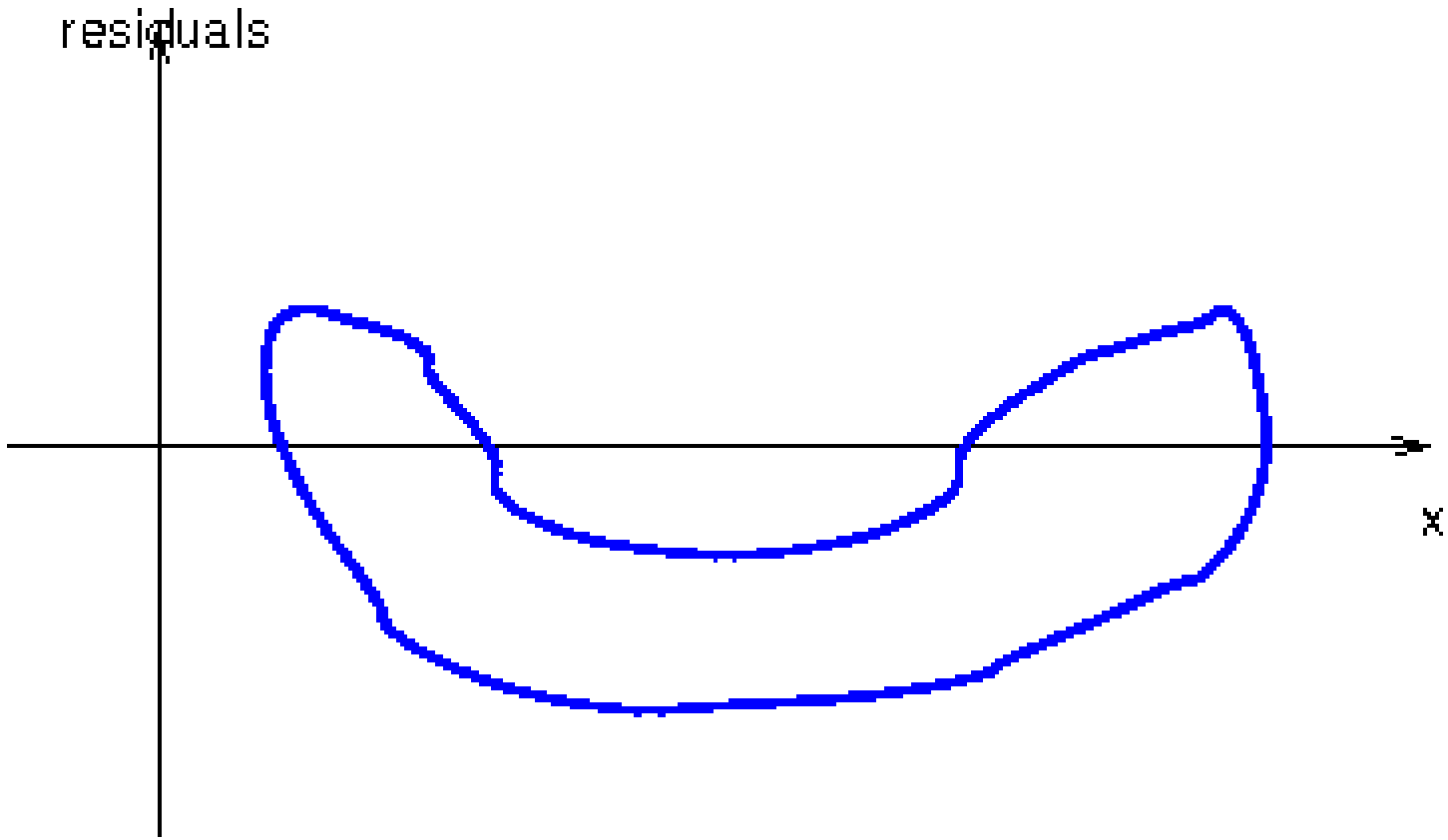


Heteroskedastic data

Fact

RMS error formula is not a good error estimate for heteroskedastic data.

Nonlinear relationship



Nonlinear relationship

- Don't use the regression line!