USP

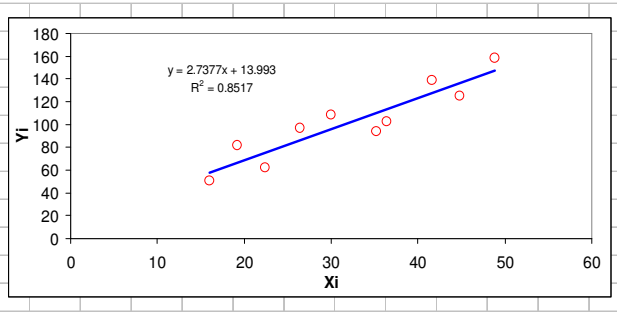
Biostatistics II

Model parameter estimations: Confronting models with measurements

Fitting models to data

- We all know how to do this

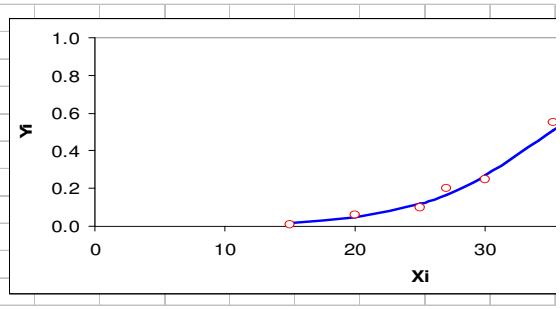
i	Xi	Yi
1	16.0	50.79
2	19.2	81.46
3	22.4	62.26
4	26.4	96.57
5	30.0	108.28
6	35.2	94.05
7	36.4	102.85
8	41.6	138.61
9	44.8	124.75
10	48.8	158.57



$$Y = a + bX$$

- But do we all know how to do this?

i	Xi	Yi
1	15.0	0.02
2	20.0	0.05
3	25.0	0.12
4	27.0	0.17
5	30.0	0.27
6	35.2	0.51
7	37.5	0.62
8	41.6	0.79
9	42.0	0.80
10	45.0	0.88



$$Y = \frac{1}{1 + e^{-k(X-b)}}$$

Model fitting procedure

- Three essential requirements:
  - Observations from a population
  - A formal predictive model with parameters to be estimated
  - A criterion to judge the goodness of fit of the model to the observations for any combination of parameter values. The criterion is often called an objective function.

$$Y_i = \hat{Y}_i + \varepsilon_i$$

    - Y_i value of observation i
    - $\hat{Y}_i$ predicted value of observation i, "the mathematical model"
    - ε_i the residual of observation i, this value is used as to calculate some criterion to judge the goodness of fit
- Parameter estimation:
  - Statistical analysis of observations (variables) where the parameters of a certain model are estimated such that the measurements are as close as possible to the predicted value given certain criteria for goodness of fit.

Goodness of fit

Observed value = Predicted value + residual

- Predicted value: Based on some formal mathematical model

Note synonyms:

- observed value = measurement
- predicted value = fitted value = expected value
- residual error = deviation = random error = residual = error = noise

Observed value = Predicted value + ε_i

$$Y_i = \hat{Y}_i + \varepsilon_i$$

- Each residual is added to the predicted value
- i stands here for a certain observation, i = 1,2,3, …n

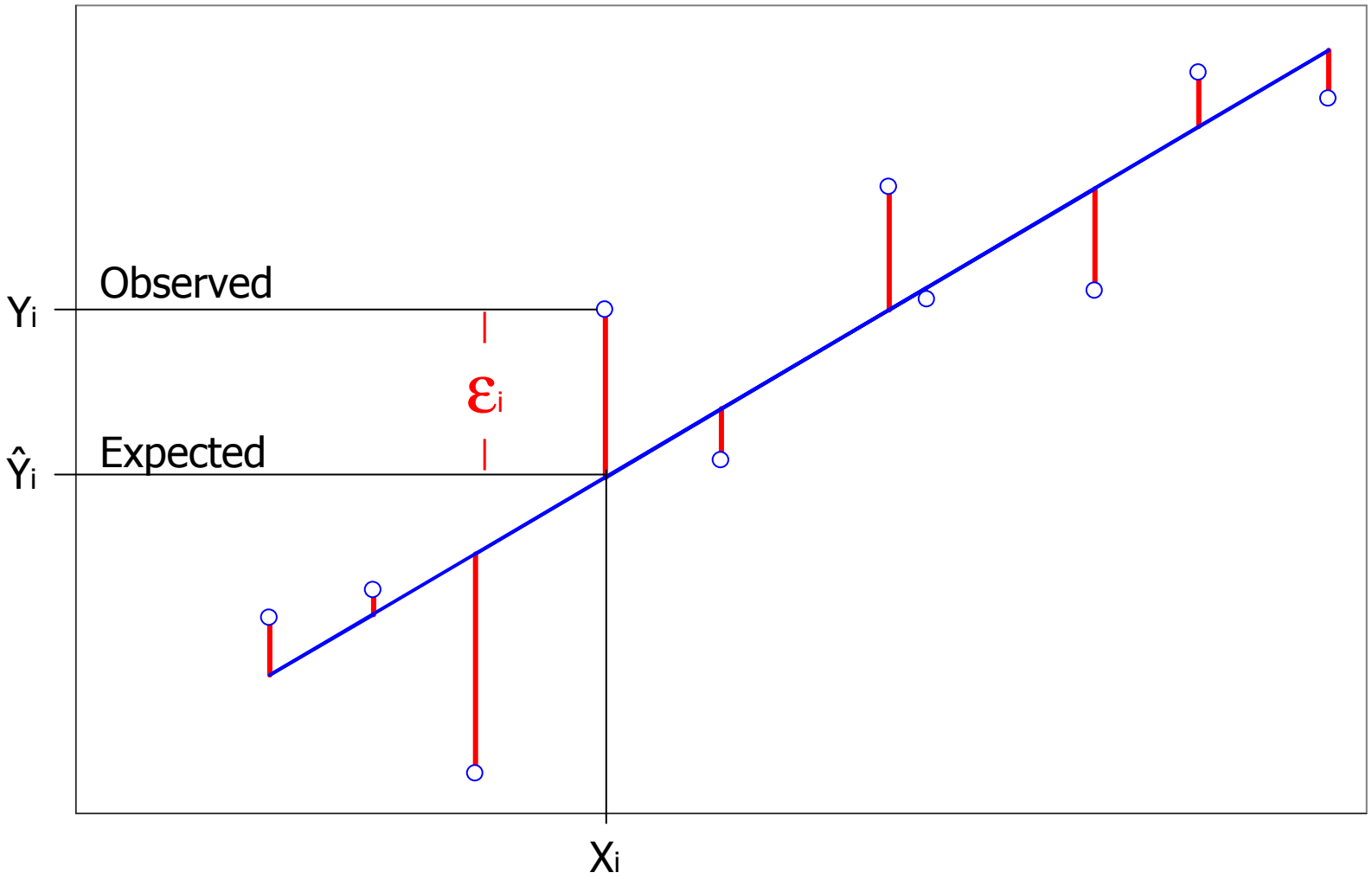
ε_i = Observed value – Predicted value

$$\varepsilon_i = Y_i - \hat{Y}_i$$

- Since the residual is a measure of distance of the prediction from that of the observed, it is an obvious candidate for measure of goodness of fit

A visual representation of residuals

$$\varepsilon_i = \left(\hat{Y}_i - Y_i\right)$$



Goodness of fit: Sum of squared residuals

ε_i = Observed value – Predicted value

- The deviations are both positive and negative
- We can thus not minimize the sum of the difference.

- Squaring the difference solves the problem of negative deviations

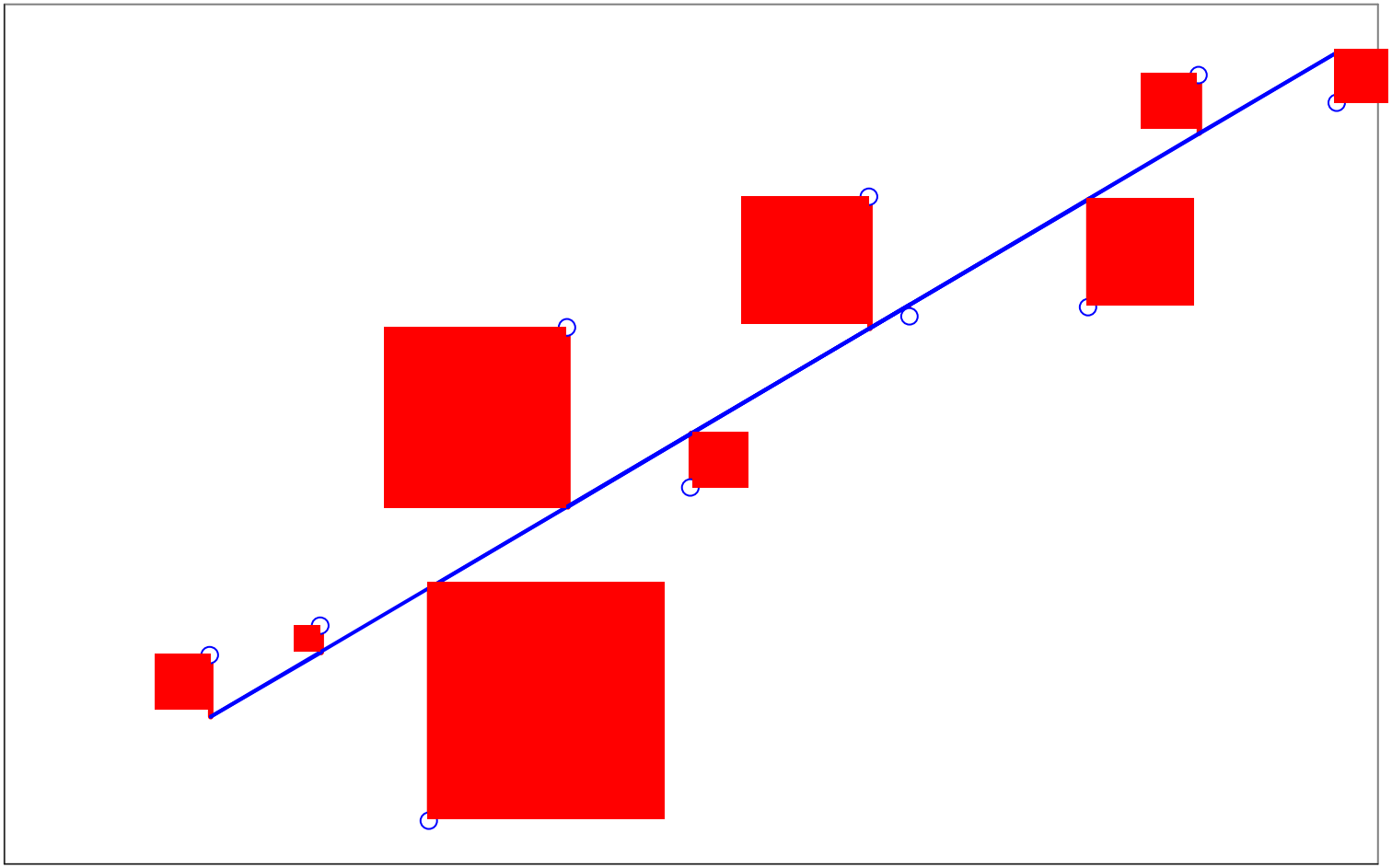
$$SS = \Sigma\varepsilon_i^2 = \Sigma(\text{Observed} - \text{Predicted})^2$$

- The criterion in the model fitting is to minimize SS

- There are 2 major assumption when using the sums of squares as the criterion of fit; The residuals are:
  - normally distributed about the predicted variable
  - with equal variance ($\sigma^{2)}$ for all values of the observed variable.

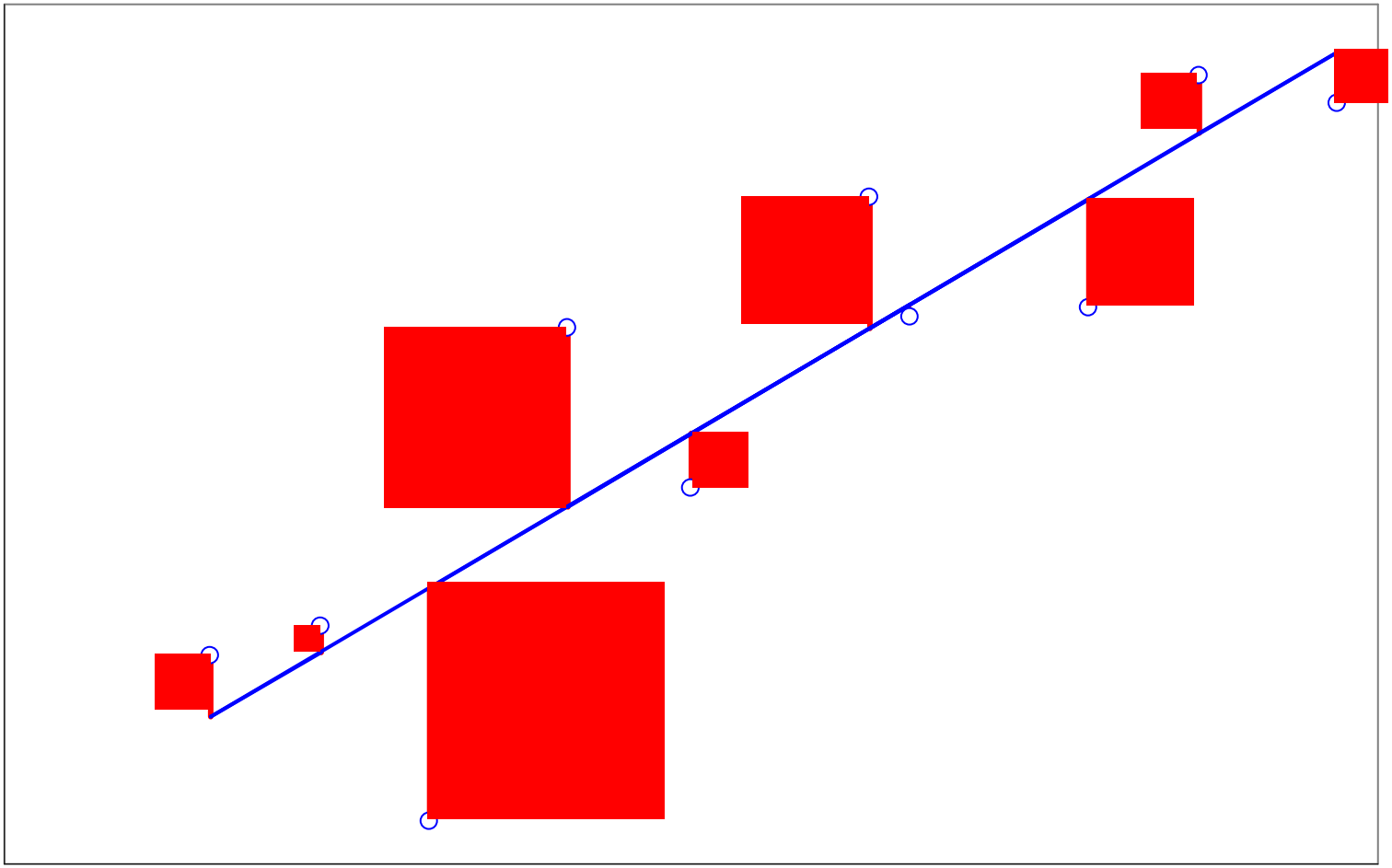
The squared residuals

$$\varepsilon_i^2 = \left(\hat{Y}_i - Y_i\right)^2$$



The sum of the squared residuals

$$\sum \varepsilon_i^2 = \sum \left(\hat{Y}_i - Y_i \right)^2 =$$



The principle behind the criterion

- Weather the model is simple or complex the principal of the criterion for the goodness of fit for the least square is always the same, i.e. minimize:
  - SS = Σ (Observed – Predicted)^2
- The only complexity is the algorithm used to obtain the best parameter estimates that describe the predicted value
- With computers it easy to search numerically for values of the parameters to find the ones that fulfill the condition of minimum sums of squares
  - Grid search: Try different values for the model parameter and calculate SS for each case. The condition is still the same: Search the value for the parameters in the model that give the lowest SS.
  - Inbuilt minimization routines: Most statistical programs have these routines. They are for all practical purposes "black boxes", how it is done is not important, the principal understanding is the issue.

In Excel the black box is called Solver

A linear model

- In mathematical notation we have:

  Observed = Predicted + residuals

  $Y_i = \hat{Y}_i + \varepsilon_i$

  $Y_i = a + b * X_i + \varepsilon_i$

  $\text{No. Eggs}_i = a + b * \text{Body weight}_i + \text{residual}$

- Thus for this model the goodness of fit is:

  $SS = \Sigma (\text{Observed}_i - \text{Predicted}_i)^2$

  $= \Sigma (Y_i - \hat{Y}_i)^2$

  $= \Sigma (Y_i - [a + b * X_i])^2$

  $= \Sigma (\text{No. Eggs}_i - [a + b * \text{Weight}_i])^2$

- Different values of the parameters a and b result in different values of SS. The objective is to find the combination of a and b that give the lowest SS value.
- SS: Sum of Squares

The multiple linear model

- In mathematical notation we have:

  Observed = Predicted + residuals

  $$Y_i = \hat{Y}_i + \varepsilon_i$$

  $$Y_i = a + b * X_i + c\, Z_i + \varepsilon_i$$

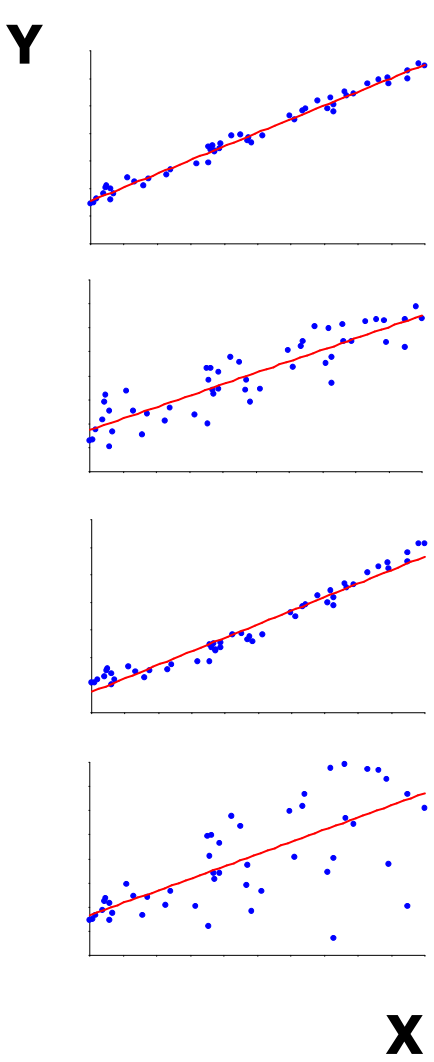
  $$\text{No. Eggs}_i = a + b * \text{Body weight}_i + \text{Temperature} + \text{residual}$$

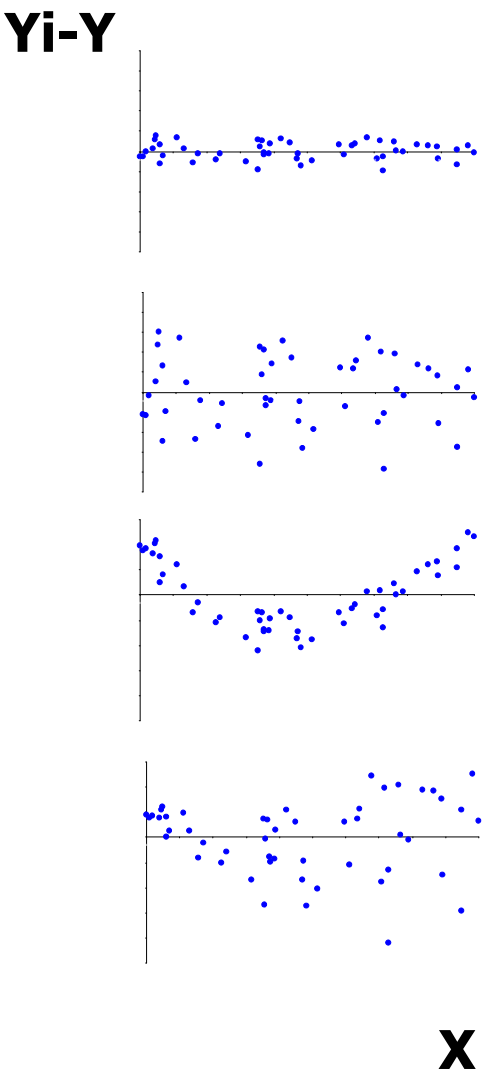
- Thus for this model the goodness of fit is:

  $$SS = \Sigma (\text{Observed}_i - \text{Predicted}_i)^2$$
  $$= \Sigma (Y_i - \hat{Y}_i)^2$$
  $$= \Sigma (Y_i - [a + b * X + c * Z_i])^2$$
  $$= \Sigma (\text{No. Eggs}_i - [a + b * \text{Weight}_i + c * \text{Temp}_i])^2$$

- Different values of the parameters a, b and c result in different values of SS. The objective to find the combination of a, b and c that give the lowest SS value.
- SS: Sum of Squares

Model violations: Residuals vs. independent value





OK, residuals random

OK, residuals random

Problem, residuals a function of x

Problem, variance increases with x