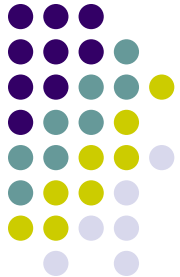
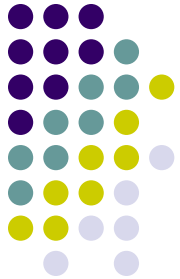
Regression Analysis



Method of **Regression Analysis** is used to **forecast** or **estimate** values of one variable (*respond variable, predicted variable*) by certain formula of one or several other variables (*descriptive variables, estimators*)

Example. There certain relation between **height** and **weight** of student. Based on data collected from *n* students, how to estimate **weight** of another student if his **height** is given?

CuuDuongThanCong.com https://fb.com/tailieudientucntt



Simple Linear Regression Model:

$$Y = a.X + b + e$$

where

* ***a*** is called ***slop*** of regression equation, informing how much the dependent variable ***Y*** grows up (or gets down) if the independent variable ***X*** increases 1 unit;

* ***b*** is called *regression constant* (intercept), showing the intersection point of regression line and vertical axis, that is the value of ***Y*** when ***X*** takes value ***0***;

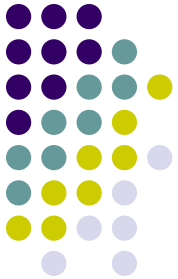
* ***e*** is *residual* of regression, indicates error of estimation at each point of observation.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Example. Concerning relation between "expenditure for buying valued items" (furniture, TV, motorbike, etc.) and "income from trading" of households in a rural area we can build up a regression equation of above linear form with "expenditure for buying valued items" as independent variable ***X*** and "income from trading" as dependent variable ***Y***. Then

- The slop ***a*** is the share for "buying valued items" in 1 VND of "income from trading"
- The intercept ***b*** allows us to know expenditure for buying valued items of given household when the household has no income from trading

CuuDuongThanCong.com https://fb.com/tailieudientucntt



Non-linear regression forms.

* Quadratic: $Y = a.X^2 + b$, Cubic: $Y = a.X^3 + b$
* Polynomial: $Y = a + b.X + c.X^2 + d.X^3 + \ldots$
* Power: $Y = e^{a.X+b} + c$
* Logarithmic: $Y = a.\log(X) + b$
* Square root: $Y = a.\sqrt{X + b} + c$
* Inverse: $Y = a/(X + b) + c$

. . .

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Non-linear regression

In many regression problems, there is no linear relation between dependent and independent variables. Then model of non-linear regression

$$Y = f(X) + e$$

(with f is a non-linear function) can be available

Example. Weight (in kg) of under 1 infant's increases systematically with age (in month) of children. However the increasing is not monotone: In first months the weight gets up more than in later months → the model of non-linear regression is more suitable than the model of linear regression.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Remark

1. For choosing a suitable regression model, it is worthy to use scatter plot to forecast a possible relation between dependent and independent variables;

2. If two regression models (e.g. linear and non-linear) give the same value of fitting then it is worthy to use the simpler model for the reason of applicability.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Estimate regression coefficients using method of least squares criterion

For linear regression model $Y = a.X + b + e$, collect a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)$$

Regression function should be of the form

$$\hat{Y}_i = \hat{a}.X_i + \hat{b}$$

$$Y_i = \hat{Y}_i + \hat{e}_i = \hat{a}.X_i + \hat{b} + \hat{e}_i$$

→ Need to estimate the regression coefficients minimizing the sum of residual (error) squares:

$$\sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (Y_i - \hat{a}X_i - \hat{b})^2 = f(\hat{a}, \hat{b}) \rightarrow \min$$

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Solution

Partial derivatives of function f vanish at the minimal point of f (sufficient condition):

$$\frac{\partial f}{\partial \hat{b}} = -\sum_{i=1}^{m} 2(Y_i - \hat{a}.X_i - \hat{b}) = 0$$

$$\frac{\partial f}{\partial \hat{a}} = -\sum_{i=1}^{m} 2(Y_i - \hat{a}.X_i - \hat{b})X_i = 0$$

Then

$$\sum_{i=1}^{m} Y_i = \hat{a}\sum_{i=1}^{m} X_i + m\hat{b}\ ;\ \sum_{i=1}^{m} Y_i X_i = \hat{a}\sum_{i=1}^{m} X_i^2 + \hat{b}\sum_{i=1}^{m} X_i$$

$$\hat{a} = \frac{\sum_{i=1}^{m} X_i Y_i - m\,\overline{X}\,\overline{Y}}{\sum_{i=1}^{m} X_i^2 - m(\overline{X})^2}\ ;\ \hat{b} = \overline{Y} - \hat{a}\overline{X}$$

CuuDuongThanCong.com https://fb.com/tailieudientucntt

2. Evaluation of model quality

Having estimated regression coefficients, we perform correspondent regression function → for each value of independent in the right hand side of regression equation we have determined value of a new variable ***Y'*** in the left hand side of the equation. This is prediction of dependent variable ***Y*** . Then

* Residual variable ***e*** equals ***Y – Y'***;

* Correlation coefficient ***R = r(Y,Y')*** between dependent variable ***Y*** and prediction variable ***Y'*** is greater than 0 and less than 1, represents the "closeness" between dependent and prediction variable. For two model with the same dependent variable and the same sample size, the model with the greater coefficient is better in forecasting, the prediction is more precise

CuuDuongThanCong.com https://fb.com/tailieudientucntt

* For simple linear regression model, R equals absolute value of correlation coefficient between dependent and independent variables $|r(X,Y)|$

* In practice, the quantity R^2 is usually used in place of R. This quantity is called ***coefficient of determination***.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

3. Evaluation regression model quality by residuals (errors) analysis

With an estimated regression model, scatter plots presenting association between residuals and dependent or independent variable can be performed for checking

a) **Homogeneity** of residuals,

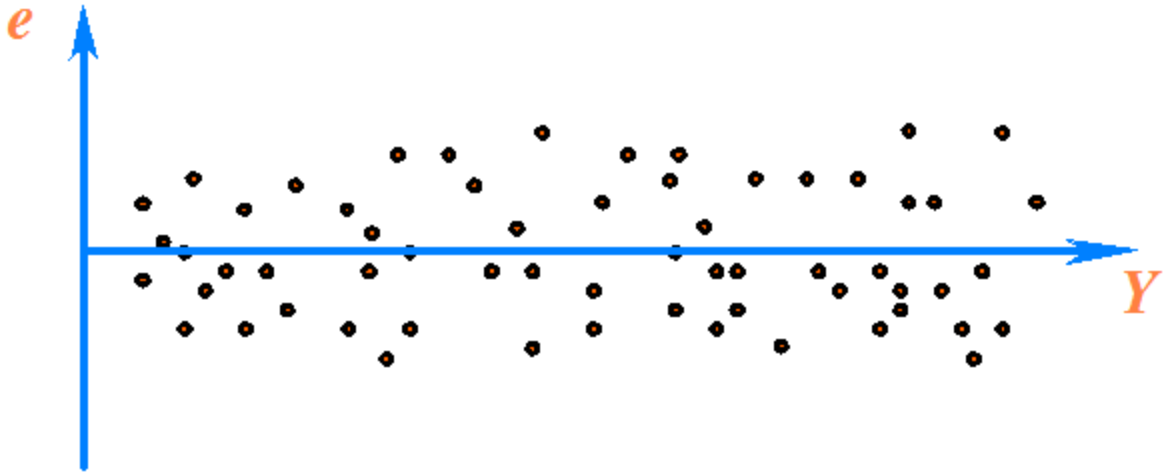
b) **Changing tendency** of residuals

And then **regulate the model** to have more suitable model

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Some possible forms of residuals distribution

Form 1:

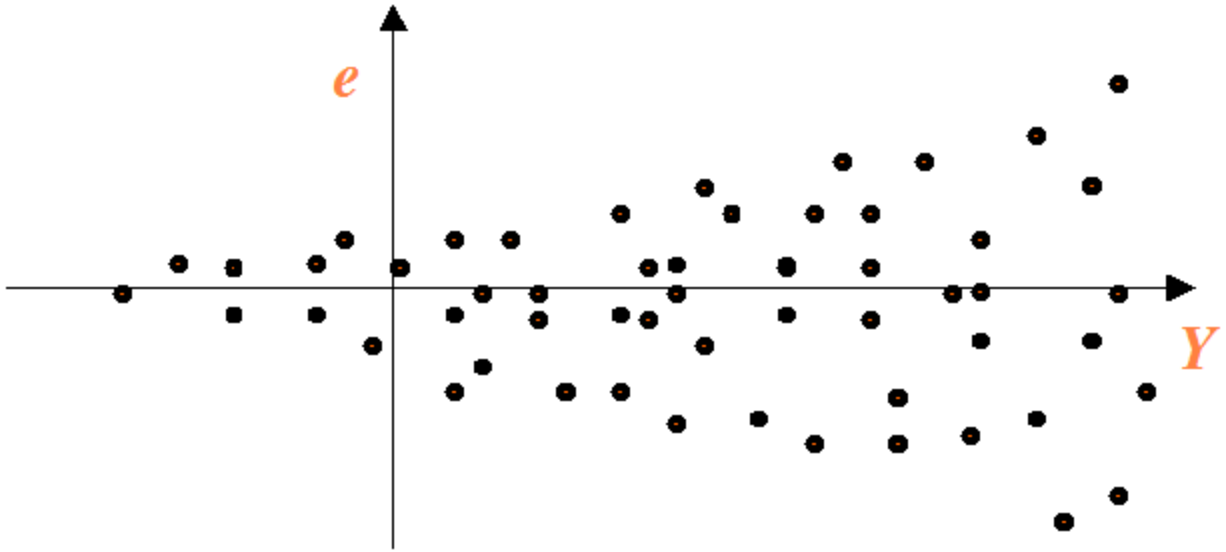


Residuals destributed in both sides and close to y axis, are almost invariant across y . Then values of variable Y have been estimated with almost the same precision.

→ The model has been correctely determined. If corelation coefficient ***R*** is still small, we can improve the model by some transformations of independent variable or adding other independent variables to the regression equation

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Form 2.

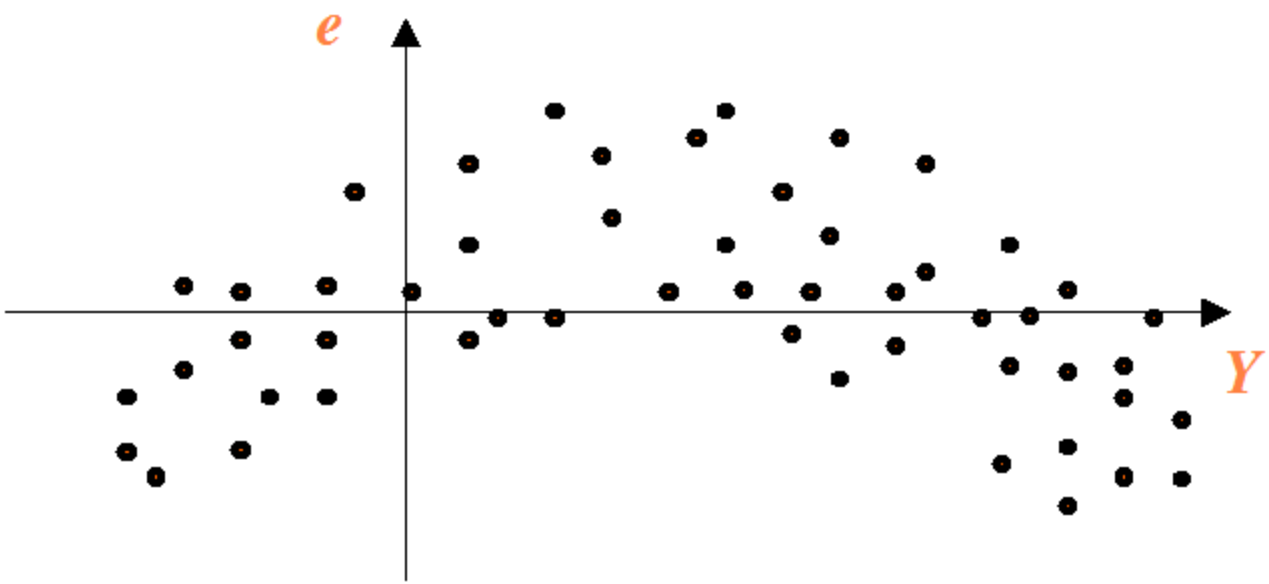


Precision of model decreases (errors are large) when y increases.

→ Transform the dependent variable Y to have a better model or use multi-level models

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Form 3



Residuals have been under estimated in certain locations and over estimated in other locations of variable y .

→ Perform plot between residuals and independent to choose other model. Non-linear models can be also considered

CuuDuongThanCong.com https://fb.com/tailieudientucntt

4. Evaluate model quality by using statistical tests

Correlation coefficient ***R(Y,Y')*** of regression model does not present completely the quality of the model

For two models with different independent variables and different sample sizes, the correlation coefficient can not provide comparison between those two models

Then suitable **tests** can be use for evaluation and choosing models.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Test 1.

Hypothesis $H: a = b = 0$

Theorem. *Consider simple linear regression model*

$$Y = a.X + b \ ,$$

with assumption of independent and Normal distributed of residuals. The variable $F(2,n-3)$ *of Fisher distribution with* $(2,n-3)$ *degrees of freedom (*n *is sample size) can be used for testing the hypothesis* H *about the vanishing of regression coefficients.*

Namely, calculate the quantity

$$s = \frac{R^2 / 2}{(1 - R^2) / (n - 2 - 1)}$$

And probability

$$p = P\{F(2,n-3) > s\}$$

Then compare p with significance level ***alpha*** to decide accept or reject the hypothesis H .

CuuDuongThanCong.com https://fb.com/tailieudientucntt

* If ***p > alpha*** → accept hypothesis ***H*** , conclude regression coefficients equal 0. Then independent variable has no influence on regression model, there is no association between that variable and dependent variable → The model is not correct, it need to find other models

* If ***p <= alpha*** → reject hypothesis ***H*** , confirm at least one of regression differs from 0 and the model is good fitted

Note. The probability ***p*** can be used for choosing model. Among two simple regression model, that with smaller probability ***p*** should be better.

CuuDuongThanCong.com https://fb.com/tailieudientucntt

Test 2.

Hypothesis Hb : b = 0
Hypothesis Ha : a = 0

The above tests can be proceeded by using a variable ***T(n-1)*** of Student distribution with ***(n-1)*** degrees of freedom (***n*** is number of observations in the regression sample)

For coefficient ***a*** we use the statistic (the procedure for ***b*** is similar)

$$t_a = \frac{\hat{a}}{se(\hat{a})}$$

to calculate probability

$$p = P\{ |T(n-1)| > |t_a| \}$$

and compare the probability with significance level ***alpha***:
- If ***p > alpha*** → accept hypothesis ***Ha*** ,
- If ***p <= alpha*** → reject hypothesis ***Ha***

CuuDuongThanCong.com https://fb.com/tailieudientucntt