Choosing independent variables

- The main idea of variables selection is to reduce the number of independent variables.
- The goal is to identify the independent variables that are decently correlated with dependent variable and possibly not correlated among themselves.
- Otherwise the independent variables might be of linear relationship which may seriously damage the model.

Choosing independent variables

Three popular methods of choosing independent variables are:

- Hellwig's method
- Graphs analysis method
- Correlation matrix method.

Hellwig's method

Three steps:

1. *Number of combinations:* 2^m-1
2. *Individual capacity of every independent variable in the combination:*

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

3. *Integral capacity of information for every combination:*

$$H_k = \sum h_{kj}$$

Hellwig's method

1. Number of combinations

In Hellwig's method the number of combinations is provided by the formula $2^m - 1$ where *m* is the number of independent variables.

Hellwig's method

2. Individual capacity of each independent variable in the combination is given by the formula:

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

where:

h_{kj} – individual capacity of information for *j*-th variable in *k*-th combination

Hellwig's method

2. Individual capacity of each independent variable in the combination is given by the formula:

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

where:

r_{0j} – correlation coefficient between *j*-th variable (independent) and dependent variable

Hellwig's method

2. Individual capacity of each independent variable in the combination is given by the formula:

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

where:

r_{ij} – correlation coefficient between *i*-th and *j*-th variable (both independent)

Hellwig's method

2. Individual capacity of each independent variable in the combination is given by the formula:

$$h_{kj} = \frac{r_{0j}^2}{\sum_{i \in I_k} |r_{ij}|}$$

where:

I_k – the set of numbers of variables in *k*-th combination

Hellwig's method

3. Integral capacity of information for every combination

The next step is to calculate H_k – integral capacity of information for each combination as the sum of individual capacities of information within each combination:

$$H_k = \sum h_{kj}$$

Hellwig's method

- *Q: HOW TO CHOOSE INDEPENDENT VARIABLES?*
- A: LOOK AT INTEGRAL CAPACITIES OF INFORMATION. THE GREATEST Hk MEANS THAT VARIABLES FROM THIS COMBINATION SHOULD BE INCLUDED IN THE MODEL.

Example

- Let’s choose independent variables, using Hellwig's method.

Y	X1	X2	X3
12	1	5	8
14	1	6	7
17	3	6	6
20	2	8	5
25	4	6	6
30	4	9	5
36	6	9	5

Example

- First we need to have vector and matrix of correlation coefficients.
  - ❑ Correlation coefficients between every independent variable X1, X2 and X3 and dependent variable Y are provided in vector R_0.

	0,9467	r_{yx1} (r_{01})
R_0	0,8374	r_{yx2} (r_{02})
	-0,7895	r_{yx3} (r_{03})

Example

- First we need to have vector and matrix of correlation coefficients.
- ❑ Correlation matrix R includes correlation coefficients between independent variables.

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

	r_{x1x1} (r_{11})	r_{x1x2} (r_{12})	r_{x1x3} (r_{13})
R	r_{x2x1} (r_{21})	r_{x2x2} (r_{22})	r_{x2x3} (r_{23})
	r_{x3x1} (r_{31})	r_{x3x2} (r_{32})	r_{x3x3} (r_{33})

Example

1. *Number of combinations*

We have 3 independent variables X1, X2 and X3. Thus we may have $2^m-1 = 2^3-1= 8-1=$ 7 combinations of independent variables.

$\{X_1\}$ $\{X_2\}$ $\{X_3\}$ $\{X_1, X_2\}$ $\{X_1, X_3\}$ $\{X_2, X_3\}$ $\{X_1, X_2, X_3\}$

Example

2. Individual capacity of independent variable in the combination 1

Combination 1

{X1} $$h_{11} = \frac{0{,}9467^2}{1} = 0{,}8962 = H_1$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

2. Individual capacity of independent variable in the combination 2

Combination 2

{X2} $$h_{22} = \frac{0{,}8374^2}{1} = 0{,}7013 = H_2$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

2. Individual capacity of independent variable in the combination 3

Combination 3

{X3} $$h_{33} = \frac{(-0,7895)^2}{1} = 0,6234 = H_3$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

2. Individual capacity of every independent variable in the combination 4

Combination 4

{X1, X2}

$$h_{41} = \frac{0{,}9467^2}{1+0{,}671} = 0{,}5364; \quad h_{42} = \frac{0{,}8374^2}{1+0{,}671} = 0{,}4197$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

2. Individual capacity of independent variables in the combination 5

Combination 5

{X1, X3}

$$h_{51} = \frac{0{,}9467^{\,2}}{1 + |-0{,}712|} = 0{,}5236; \quad h_{53} = \frac{(-0{,}7895)^2}{1 + |-0{,}712|} = 0{,}3642$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

2. Individual capacity of every independent variables in the combination 6

Combination 6

{X2, X3}

$$h_{62} = \frac{0{,}8374^{\,2}}{1 + |-0{,}884|} = 0{,}3723; \quad h_{63} = \frac{(-0{,}7895)^2}{1 + |-0{,}884|} = 0{,}3309$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

Combination 7

{X1, X2, X3}

$$h_{71} = \frac{0{,}9467^2}{1+0{,}671+\left|-0{,}712\right|} = 0{,}3762; \quad h_{72}\ \frac{0{,}8374^2}{0{,}671+1+\left|-0{,}884\right|} = 0{,}2745$$

$$h_{73} = \frac{(-0{,}7895)^2}{\left|-0{,}712\right|+\left|-0{,}884\right|+1} = 0{,}2402$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

3. Integral capacity of information for each combination

Combination number	Independent variables for combination	Individual capacity of information **h**		Integral capacity of information **H**	
1	**{X_1}**	h_{11}	0,8962	H_1	**0,89621**
2	**{X_2}**	h_{22}	0,7013	H_2	**0,7013**
3	**{X_3}**	h_{33}	0,6234	H_3	**0,62338**
4	**{X_1, X_2}**	h_{41}	0,5364	**H_4**	**0,95612**
		h_{42}	0,4197		
5	**{X_1, X_3}**	h_{51}	0,5236	H_5	**0,88786**
		h_{53}	0,3642		
6	**{X_2, X_3}**	h_{62}	0,3723	H_6	**0,70316**
		h_{63}	0,3309		
7	**{X_1, X_2, X_3}**	h_{71}	0,3762	H_7	**0,89089**
		h_{72}	0,2745		
		h_{73}	0,2402		

The greatest integral capacity is for combination C4. Independent variables - X1, X2 - will be included in model.

Graph analysis method

Three steps

1. *Calculating r**
2. *Modification of correlation matrix*
3. *Drawing the graph*

Graph analysis method

- *Q: HOW TO CHOOSE INDEPENDENT VARIABLES?*

- A: LOOK AT THE GRAPHS. THE NUMBER OF GROUPS MEANS THE NUMBER OF VARIABLES INCLUDED IN THE MODEL. IF THERE'S SEPARATED (ISOLATED) VARIABLE, YOU SHOULD INCLUDE IT IN THE MODEL. FROM EACH GROUP, THE VARIABLE WITH THE GREATEST NUMBER OF LINKS SHOULD BE INCLUDED IN MODEL. IF THERE'S TWO VARIABLES WITH THE GREATEST NUMBER OF LINKS, YOU SHOULD TAKE THE VARIABLE WHICH IS MORE STRONGLY CORRELATED WITH DEPENDENT VARIABLE.

Graph analysis method

1. *Calculating r**

We start with calculating critical value of r* using the formula:

$$r^* = \sqrt{\frac{t_\alpha^2}{n-2+t_\alpha^2}}$$

where t_α is provided in the table of t-Student distribution at the significance level *α* and the degrees of freedom *n-2* (sometimes r* can be given, so there's no need to calculate it).

Graph analysis method

2. Modification of correlation matrix

The correlation coefficients for which $\left|r_{ij}\right| \leq r^*$ are statistically irrelevant and we replace them with nulls in correlation matrix.

3. Drawing the graph

Using modified correlation matrix we draw the graphs with bulbs representing the variables and the links representing correlation coefficients of statistical significance.

Example

Let's have an example *(the same one as for Hellwig's method, n=7)*

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

1. Calculating r (n=7, $t_{\alpha,\, n-2}=t_{0,05,5}=2,571$)*

$$r^* = \sqrt{\frac{t_\alpha^2}{n-2+t_\alpha^2}} = \sqrt{\frac{2,571^2}{5+2,571^2}} = \sqrt{\frac{6,61}{11,61}} = \sqrt{0,569337} = 0,7545$$

Example

2. *Modification of correlation matrix*

$$\left|r^{*}\right| \leq 0{,}7545$$

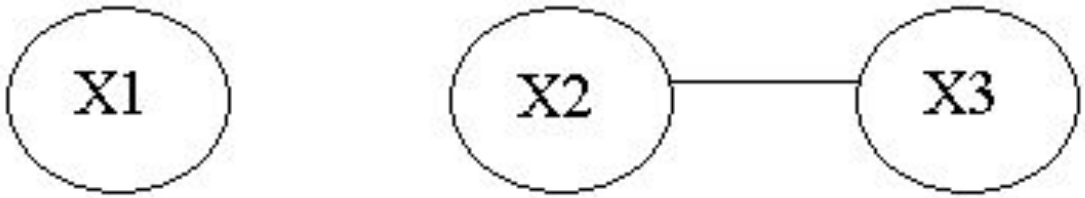
	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

	1	0	0
R*	0	1	-0,884
	0	-0,884	1

Example

3. Drawing the graph

	1	0	0
R*	0	1	-0,884
	0	-0,884	1



Conclusion: Model will consist of X1 (as isolated variable) and x2 (cause is more strongly correlated with dependent variable – you may check it in R_0 vector).

	0,9467
R_0	0,8374
	-0,7895

Correlation matrix method

1. *Calculate r**

We start with calculating critical value of r* using the formula:

$$r^* = \sqrt{\frac{t_\alpha^2}{n-2+t_\alpha^2}}$$

where t_α is provided in the table of t-Student distribution at the significance level *α* and the degrees of freedom *n-2* (sometimes r* can be given, so there's no need to calculate it).

2. To eliminate X_i variables weakly correlated with Y

$$\left|r_{ij}\right| \leq r^*$$

3. To choose X_s where $\left|r_s\right| = \max\left\{\left|r_i\right|\right\}$

[X_s is the best source of information]

4. To eliminate X_i variables strongly correlated with Xs

$$\left|r_{si}\right| > r^*$$

Example

Let's have an example *(the same one as for Hellwig's method and graph analysis metod, n=7)*

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

Example

1. Calculating r (n=7, $t_{\alpha,\, n-2}=t_{0,05,5}=2,571$)*

$$r^* = \sqrt{\frac{t_\alpha^2}{n-2+t_\alpha^2}} = \sqrt{\frac{2,571^2}{5+2,571^2}} = \sqrt{\frac{6,61}{11,61}} = \sqrt{0,569337} = 0,7545$$

2. To eliminate X_i variables weakly correlated with Y

$$\left|r_{ij}\right| \le r^*$$

$$r^* = 0{,}7545$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

None of the variables will be eliminated

3. To choose X_s where $\left|r_s\right| = \max\left\{\left|r_i\right|\right\}$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

4. To eliminate X_i variables strongly correlated with Xs

$$|r_{si}| > r^* \qquad r^* = 0{,}7545$$

	0,9467
R_0	0,8374
	-0,7895

	1	0,671	-0,712
R	0,671	1	-0,884
	-0,712	-0,884	1

None of the variables will be eliminated.

X1, X2, X3 will be included in model.

In this example level of significane can be changed – this will give us different results (you may check it if you want).

DON’T EXPECT TO GET THE SAME RESULTS FROM THESE THREE METHODS…