

Regresszió

Pusztaházi Luca Sára
Ország Anna

Budapesti Műszaki és Gazdaságtudományi Egyetem
Matematika Intézet
Sztochasztika Tanszék



2017

Tartalomjegyzék

Elméleti háttér

Módszertan

Lineáris regresszió

Használat

- A regresszióanalízis során két vagy több véletlen változó között fennálló kapcsolatot modellezünk.
- Lehet lineáris és nemlineáris a regresszió a modell tulajdonságai alapján.
- A regresszió feladata az $Y = l(X_1, X_2, \dots, X_n)$ függvény meghatározása.
- Y : függő (dependent) változó
- X_1, X_2, \dots, X_n : magyarázó (predictor) változók

- Cél: az Y valószínűségi változó közelítése az $l(X_1, X_2, \dots, X_n)$ függvénnyel legkisebb négyzetes értelemben.
- Azaz az $\mathbb{E}(Y - l(X_1, X_2, \dots, X_n))^2$ várhatóérték minimalizálása.
- Belátható, hogy ezt az $\mathbb{E}(Y|X_1, X_2, \dots, X_n)$ feltételes várhatóérték elégíti ki.
- Ezt az együttes sűrűségfüggvényből $f(Y, X_1, X_2, \dots, X_n)$ meg lehetne határozni, de általában ez nem adott, csak egy statisztikai mintánk van a függő és független változókra.

- Lineáris regresszió

$$Y = a_1X_1 + a_2X_2 + \dots + a_pX_p + b$$

- Multiplikatív modell

$$Y = bX_1^{a_1} X_2^{a_2} \dots X_p^{a_p}$$

- Polinomiális regresszió

$$Y = a_1X + a_2X^2 + \dots + a_pX^p + b$$

1. Adatok vizsgálata

- A független változók normális eloszlásúak?
- Vannak extrém értékű, kihagyható megfigyelések?
- Pontdiagram alapján az $y - x$ párok lineáris kapcsolata fennáll?
- A független változók közötti páronkénti korrelációk gyengék? Ha nem, akkor szakmai vagy statisztikai szempontok alapján válogatjuk ki a magyarázó változókat?

2. Az illesztés menete, a változók közötti szelekció végrehajtása
- Melyek a statisztikai értelemben legerősebb magyarázó erővel bíró változók?
 - Létezik-e lineáris modell vagy minden becsült együttható nullának tekinthető?
 - Milyen tesztekkel és hogyan minősíthető a modell egésze?

3. A magyarázó változók közötti kapcsolatrendszer megfelelő?
- Milyen mutatókra támaszkodhatunk annak mérésekor, hogy túlzott multikollinearitás fellépett-e?
 - Mely változók elhagyásával küszöbölhető ki a multikollinearitás?

4. Modell diagnosztika, hibatagok viselkedése, kiugró pontok kezelése

- Megfelelő magyarázó erejű modellt kaptunk?
- A hibatagok normális eloszlásúak?
- A hibatagok szórása azonos-e, nem lépett fel heteroszkedaszticitás?
- Vannak nagyon erős hatást gyakorló megfigyelések a mintában? Ezek elhagyása indokolt?

- Az $Y = a_1X_1 + a_2X_2 + \dots + a_pX_p + b$ egyenletben az együtthatók és a konstans meghatározása
- Azaz:

$$\min_{a_1, \dots, a_p, b} g(a_1, \dots, a_p, b) = \min_{a_1, \dots, a_p, b} \mathbb{E}(Y - (a_1X_1 + \dots + a_pX_p + b))^2$$

- b szerinti derivált:

$$\frac{\partial g}{\partial b} = -2\mathbb{E}(Y - (a_1X_1 + \dots + a_pX_p + b)) = 0$$

- Kifejezve b -t:

$$b = \mathbb{E}Y - a_1\mathbb{E}X_1 - \dots - a_p\mathbb{E}X_p$$

- a_i ($i = 1, \dots, p$) szerinti derivált:

$$\frac{\partial g}{\partial a_i} = -2\mathbb{E}((Y - (a_1X_1 + \dots + a_pX_p + b)) X_i) = 0$$

- b -t behelyettesítve az a_i szerinti deriváltba:

$$\begin{aligned} 0 &= -2\mathbb{E}((Y - (a_1X_1 + \dots + a_pX_p + b)) X_i) \\ &= -2\mathbb{E}\left(\left((Y - \mathbb{E}Y) - \sum_{j=1}^p a_j(X_j - \mathbb{E}X_j)\right) X_i\right) \\ &= -2(\mathbb{E}(YX_i) - \mathbb{E}Y\mathbb{E}X_i) + \\ &\quad + 2\sum_{j=1}^p a_j(\mathbb{E}(X_iX_j) - \mathbb{E}X_i\mathbb{E}X_j) \\ &= -2\text{Cov}(Y, X_i) + 2\sum_{j=1}^p a_j\text{Cov}(X_i, X_j), \quad i = 1, \dots, p. \end{aligned}$$

Lineáris regresszió elméleti megoldása III.

- Jelölés: $\mathbf{C} := \text{Var}(\underline{X})$, $\underline{d} := \text{Cov}(Y, \underline{X})$
- Ebből a következő egyenletrendszert kapjuk:

$$\mathbf{C}\underline{a} = \underline{d},$$

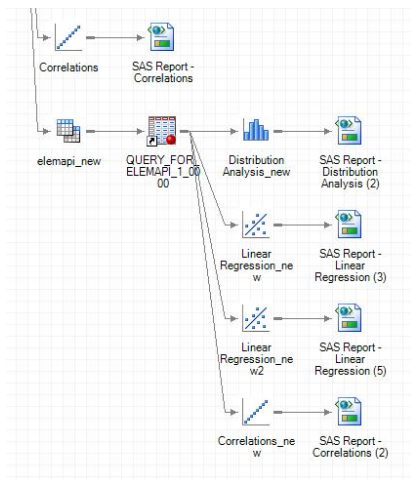
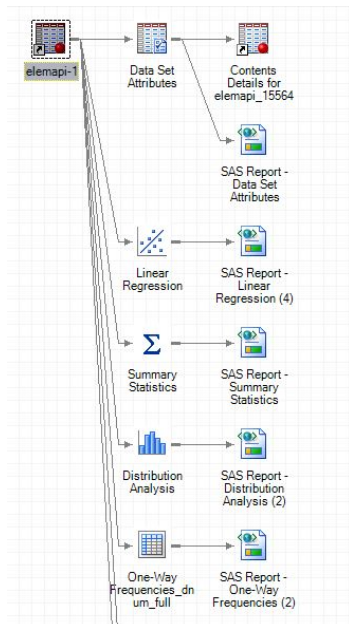
- melynek egyértelmű megoldása, ha \mathbf{C} reguláris:

$$\underline{a} = \mathbf{C}^{-1}\underline{d}.$$

- A második derivált vizsgálatával belátható, hogy ez valóban minimum.

- California Department of Education's Academic Performance Index (API) 1999-2000 Growth
- *api00*: API 2000
- *acs_k3*: átlagos osztálylétszám (k-3)
- *meals*: ingyenes vagy kedvezményes árú ebédprogramban résztvevő diákok százalékos aránya
- *full*: kitűnő ajánlással rendelkező tanárok százalékos aránya
- A 2000-es teljesítmény indexhez szeretnénk lineáris regressziót készíteni a másik három paraméter segítségével.
- *dnum*: körzetszám

Folyamatábra

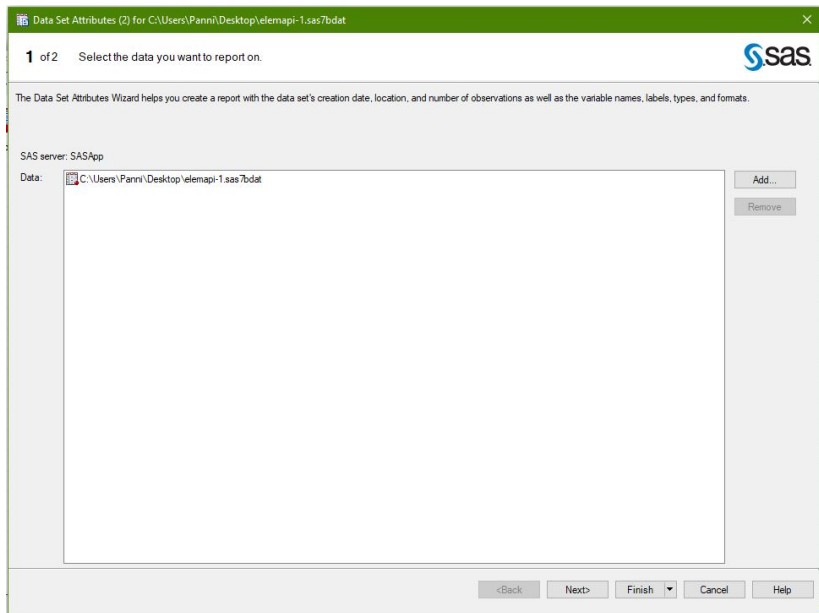


- Tasks/Data/Data Set Attributes
- Tasks/Regression/Linear Regression
- Tasks/Describe/Summary Statistics
- Tasks/Describe/Distribution Analysis
- Tasks/Describe/One-Way Frequencies
- Tasks/Multivariate/Correlation
- Tasks/Data/Query Builder

1. lépés – Ismerkedés az adatokkal

- Data Set Attributes: kapunk egy összefoglalót az adatokról, illetve megnézhetjük a változók típusát és leírását

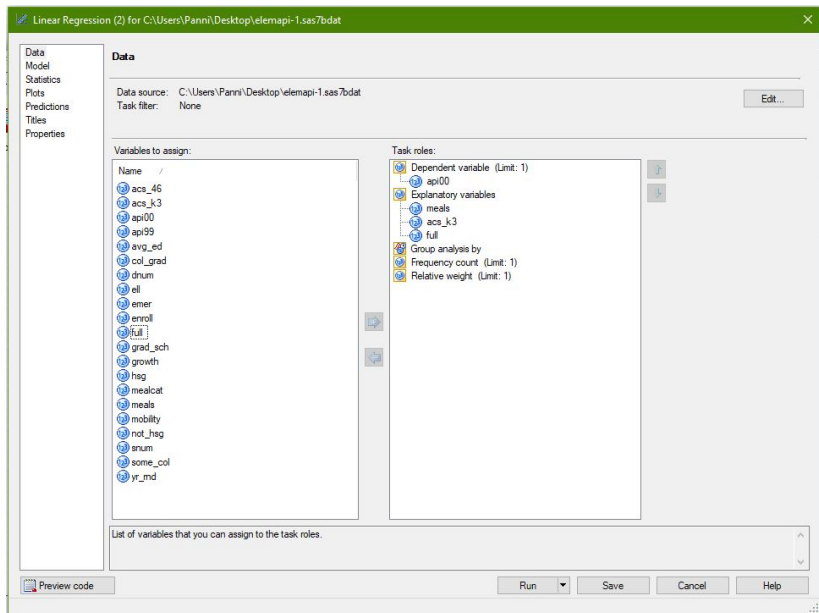
Data Set Attributes



2. lépés – Lineáris regresszió a (talán rossz) adatokra

- H_0 : a függő változó nem függ a magyarázó változóktól, azaz a regressziós egyenes meredeksége nulla.
- A futtatás után észrevehetjük, hogy az *acs_k3* és a *full* változók esetén nem utasíthatjuk el a null-hipotézist, azaz ezekkel nem lehet szignifikánsan közelíteni a teljesítmény indexet.
- A regressziós ábrán megfigyelhetjük, hogy érdekes a pontok eloszlása → nézzük meg alaposabban az adatokat!

Linear Regression



The screenshot shows the SAS Linear Regression dialog box. The title bar reads "Linear Regression (2) for C:\Users\Panni\Desktop\elemapi-1.sas7bdat". The "Data" tab is selected in the left-hand menu. The "Data" section shows the "Data source" as "C:\Users\Panni\Desktop\elemapi-1.sas7bdat" and the "Task filter" as "None". Below this, there are two main panels: "Variables to assign:" and "Task roles:". The "Variables to assign:" panel lists 25 variables: acs_46, acs_k3, api00, api99, avg_ed, col_grad, dnum, ell, emer, enroll, full, grad_sch, growth, hsg, mealcat, meals, mobility, not_hsg, snum, some_col, and yr_md. The "Task roles:" panel shows a hierarchical structure: "Dependent variable (Limit: 1)" with "api00" assigned; "Explanatory variables" with "meals", "acs_k3", and "full" assigned; "Group analysis by" with "Frequency count (Limit: 1)" and "Relative weight (Limit: 1)" assigned. At the bottom, there are buttons for "Preview code", "Run", "Save", "Cancel", and "Help".

Linear Regression (2) for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Data

Data source: C:\Users\Panni\Desktop\elemapi-1.sas7bdat Edit...

Task filter: None

Variables to assign:

Name /

- acs_46
- acs_k3
- api00
- api99
- avg_ed
- col_grad
- dnum
- ell
- emer
- enroll
- full
- grad_sch
- growth
- hsg
- mealcat
- meals
- mobility
- not_hsg
- snum
- some_col
- yr_md

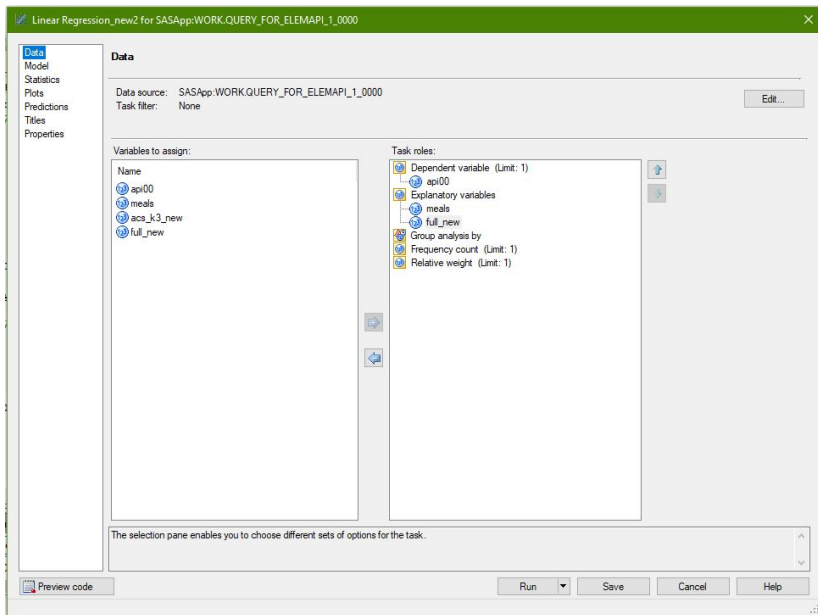
Task roles:

- Dependent variable (Limit: 1)
 - api00
- Explanatory variables
 - meals
 - acs_k3
 - full
- Group analysis by
 - Frequency count (Limit: 1)
 - Relative weight (Limit: 1)

List of variables that you can assign to the task roles.

Preview code Run Save Cancel Help

Linear Regression



The screenshot shows the SAS Linear Regression dialog box. The title bar reads "Linear Regression_new2 for SASApp:WORK.QUERY_FOR_ELEMAPI_1_0000". On the left is a navigation pane with options: Data (selected), Model, Statistics, Plots, Predictions, Titles, and Properties. The main area is titled "Data" and contains the following information:

- Data source: SASApp:WORK.QUERY_FOR_ELEMAPI_1_0000
- Task filter: None

Below this is an "Edit..." button. The main configuration area is divided into two panes:

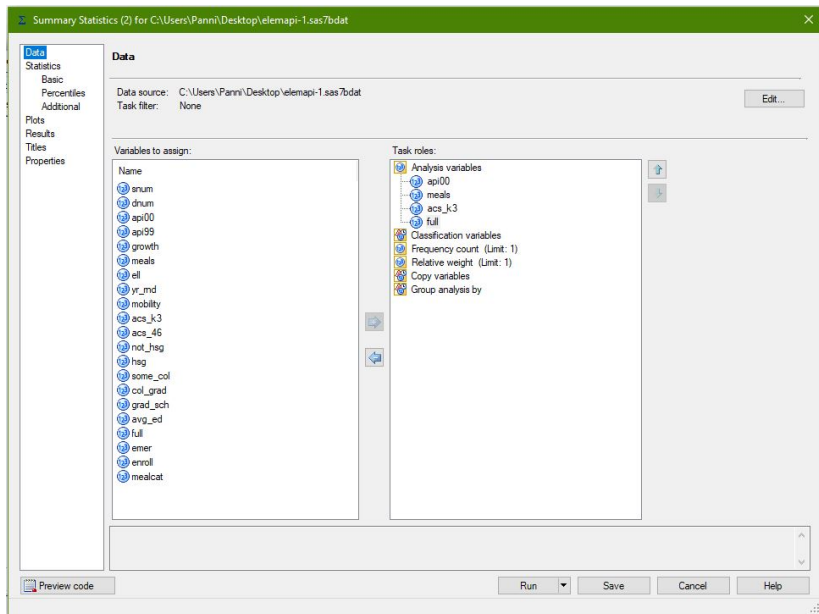
- Variables to assign:** A list of variables with selection icons: api00, meals, acs_k3_new, and full_new.
- Task roles:** A list of roles with selection icons: Dependent variable (Limit: 1) with api00 assigned; Explanatory variables with meals and full_new assigned; Group analysis by; Frequency count (Limit: 1); and Relative weight (Limit: 1).

Navigation arrows are present between the panes. At the bottom, there is a text box: "The selection pane enables you to choose different sets of options for the task." Below this are buttons for "Preview code", "Run", "Save", "Cancel", and "Help".

3. lépés – Mélyebb ismerkedés az adatokkal

- Summary Statistics: kapunk egy összefoglalást a felhasználandó változókról (min, max, átlag stb.)
- Figyeljük meg, hogy az *acs_k3* és a *full* változók esetén a minimum érték negatív, ill. nagyon kicsi – osztálylétszám és százalékos érték esetén ez nem tűnik jónak – lehet, hogy hibás az adat?
- Distribution Analysis: Figyeljük meg a fentebb kiemelt két változó hisztogramját!

Summary Statistics



The screenshot shows the SAS Summary Statistics dialog box for the file 'C:\Users\Panni\Desktop\elemapi-1.sas7bdat'. The 'Data' tab is active, showing the data source and task filter. The 'Variables to assign' list contains 25 variables, and the 'Task roles' list includes Analysis variables, Classification variables, Frequency count, Relative weight, Copy variables, and Group analysis by.

Summary Statistics (2) for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Data

Data source: C:\Users\Panni\Desktop\elemapi-1.sas7bdat
Task filter: None

Variables to assign:

Name
snum
dnum
api00
api99
growth
meals
ell
yr_md
mobility
acs_k3
acs_46
not_hsg
hsg
some_col
col_grad
grad_sch
avg_ed
full
emer
enroll
mealcat

Task roles:

- Analysis variables
 - api00
 - meals
 - acs_k3
 - full
- Classification variables
- Frequency count (Limit: 1)
- Relative weight (Limit: 1)
- Copy variables
- Group analysis by

Buttons: Preview code, Run, Save, Cancel, Help

Distribution Analysis

Distribution Analysis (2) for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Data

Data source: C:\Users\Panni\Desktop\elemapi-1.sas7bdat Edit...

Task filter: None

Variables to assign:

Name
snum
dnum
api00
api99
growth
meals
ell
yr_md
mobility
acs_k3
acs_46
not_hsg
hsg
some_col
col_grad
grad_sch
avg_ed
full
emer
enroll
mealcat

Task roles:

- Analysis variables
 - api00
 - meals
 - acs_k3
 - full
- Group analysis by
- Frequency count (Limit: 1)
- Relative weight (Limit: 1)
- Classification variables (Limit: 2)














Preview code Run Save Cancel Help

Distribution Analysis

Distribution Analysis (2) for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Plots > Appearance

Note: Insets are valid on histogram, probability and quantile-quantile plots only.

		Axis color:	Background color:	Axis width:
	<input checked="" type="checkbox"/> Histogram Plot			1
	<input type="checkbox"/> Probability Plot			1
	<input type="checkbox"/> Quantiles plot			1
	<input type="checkbox"/> Box plot			1
	<input type="checkbox"/> Text-based plots			

Produces a stem and leaf plot or bar chart (depending on the number of observations), box plot and normal probability plot. Produces a side-by-side plot if there is a by variable.

Creates a histogram and optionally superimposes density curves for continuous theoretical distributions and for kernel density estimates.

Preview code

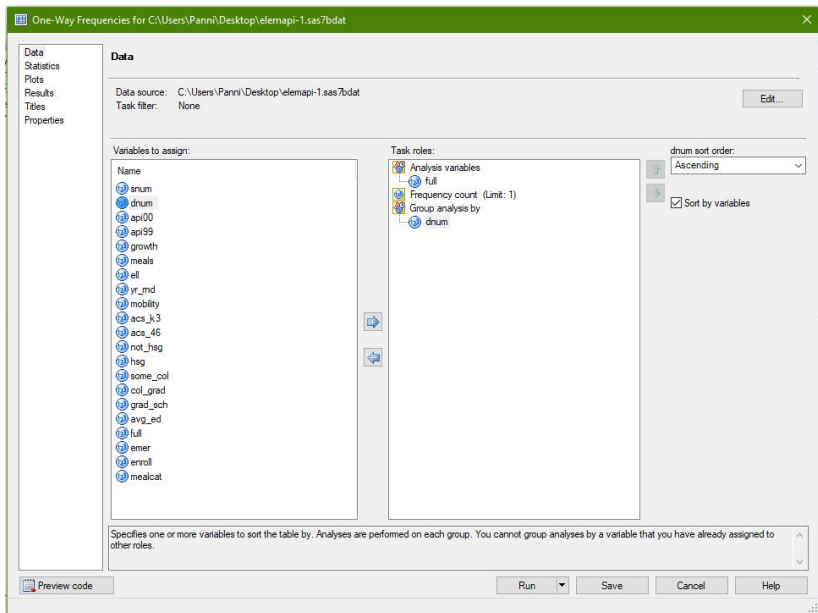
Run Save Cancel Help

The "Analysis variables" role must have at least 1 variable assigned to it.

3. lépés – Következtetés

- Az *acs_k3* változó esetén azt gondoljuk, hogy elírás történt, így később az adatsor abszolútértékével fogunk tovább dolgozni.
- A *full* változó esetén készítünk körzetszám szerint egy gyakoriságtáblát, ahol azt fogjuk látni, hogy minden körzet 1 – 100-ig terjedő skálán adta meg a tanárok %-os értékelését, kivéve a 401-es körzet – így ebben az esetben értelmezési hibáról beszélhetünk.
- Megoldás: Az 1-nél kisebb adatokat megszorozzuk 100-zal.

One-Way Frequencies



The screenshot shows the SAS One-Way Frequencies dialog box for the file 'C:\Users\Panni\Desktop\elemapi-1.sas7bdat'. The 'Data' tab is active, showing the data source and task filter. The 'Variables to assign' list includes variables like snum, dnum, api00, etc. The 'Task roles' list shows 'Analysis variables' with 'full' and 'Frequency count (Limit: 1)' assigned to 'dnum'. The 'dnum sort order' is set to 'Ascending' and 'Sort by variables' is checked. A 'Preview code' button is at the bottom left, and 'Run', 'Save', 'Cancel', and 'Help' buttons are at the bottom right.

One-Way Frequencies for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Data

Data source: C:\Users\Panni\Desktop\elemapi-1.sas7bdat
Task filter: None

Variables to assign:

Name

- snum
- dnum
- api00
- api99
- growth
- meals
- ell
- yr_md
- mobility
- acs_k3
- acs_46
- not_hsg
- hsg
- some_col
- col_grad
- grad_sch
- avg_ed
- full
- emer
- enroll
- mealcat

Task roles:

- Analysis variables
 - full
 - Frequency count (Limit: 1)
 - Group analysis by
 - dnum

dnum sort order: Ascending

Sort by variables

Specifies one or more variables to sort the table by. Analyses are performed on each group. You cannot group analyses by a variable that you have already assigned to other roles.

Preview code

Run Save Cancel Help

4. lépés – Adatsorok kijavítása

- A fentebb leírt módon létrehozunk egy új adattáblát a kijavított adatokkal.

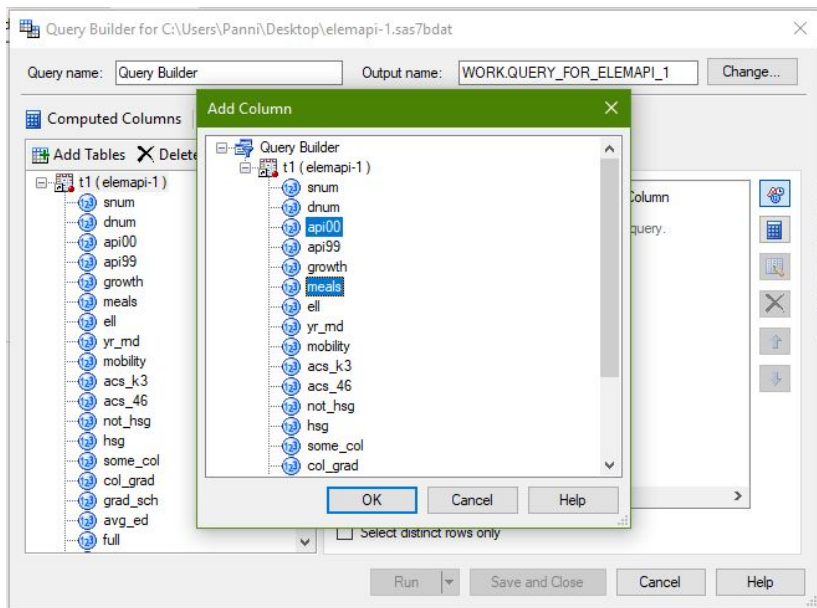
elemapi_new ▾

Input Data Code Log Output Data

Modify Task Filter and Sort Query Builder Where | Da


	api00	meals	acs_k3_new	full_new
1	693	67	16	76
2	570	92	15	79
3	546	97	17	68
4	571	90	20	87
5	478	89	18	87

Query Builder



Query Builder

New Computed Column × ×

1 of 4 Select a type 

Summarized column

Recoded column

Advanced expression

From an existing computed column

Column	Details
--------	---------

Convert to an advanced expression

<Back Next> Finish Cancel Help

Query Builder

New Computed Column

2 of 4 Build an advanced expression

sas

Enter an expression:

abs(1.acs_k3|

Home Next Back End Undo Redo Edit Favorites Validate

+ - * / ** || (x) 'x' "x" , 'abc'n

meals
ell
yr_md
mobility
acs_k3
acs_46
not_hsg
hsg
some_col
col_grad
grad_sch


The maximum number of rows to process for retrieving distinct values may be limited.

Get Values Select Values

<Back Next Finish Cancel Help

Query Builder

New Computed Column ×

3 of 4 Modify additional options 

Column Name:

Label:

Summary: Length (in bytes):

Expression:

Format:

Query Builder

New Computed Column

2 of 4 Build an advanced expression

sas

Enter an expression:

```
CASE WHEN t1.full <= 1 THEN 100*t1.full ELSE t1.full END
```

Home Next Back End Undo Redo Edit Favorites Validate

+ - * / ** || (x) ' ' " " , 'abc'n

- not_hsg
- hsg
- some_col
- col_grad
- grad_sch
- avg_ed
- full
- emer
- enroll
- mealcat

Selected Columns

The maximum number of rows to process for retrieving distinct values may be limited.

Get Values Select Values

<Back Next Finish Cancel Help

Query Builder

The screenshot shows the SAS Query Builder window titled "elemapi_new for C:\Users\Panni\Desktop\elemapi-1.sas7bdat". The window has a title bar with a close button. Below the title bar, there are two input fields: "Query name:" with the value "elemapi_new" and "Output name:" with the value "WORK.QUERY_FOR_ELEMAPI_1_001", followed by a "Change..." button. A toolbar contains icons for "Computed Columns", "Prompt Manager", "Preview", "Tools", and "Options".

Below the toolbar, there are three buttons: "Add Tables", "Delete", and "Join Tables". The main area is divided into two panes. The left pane shows a tree view of a table named "t1 (elemapi-1)" with a list of columns: snum, dnum, api00, api99, growth, meals, ell, yr_md, mobility, acs_k3, acs_46, not_hsg, hsg, some_col, col_grad, grad_sch, avg_ed, and full. The right pane is titled "Select Data" and contains a table with the following columns: "Column Name" and "Source Column".

Column Name	Source Column
api00 (api 2000)	t1.api00
meals (pct free meals)	t1.meals
acs_k3_new	Computed
full_new	Computed

Below the table in the right pane, there is a checkbox labeled "Select distinct rows only" which is currently unchecked. At the bottom of the window, there are four buttons: "Run", "Save and Close", "Cancel", and "Help".

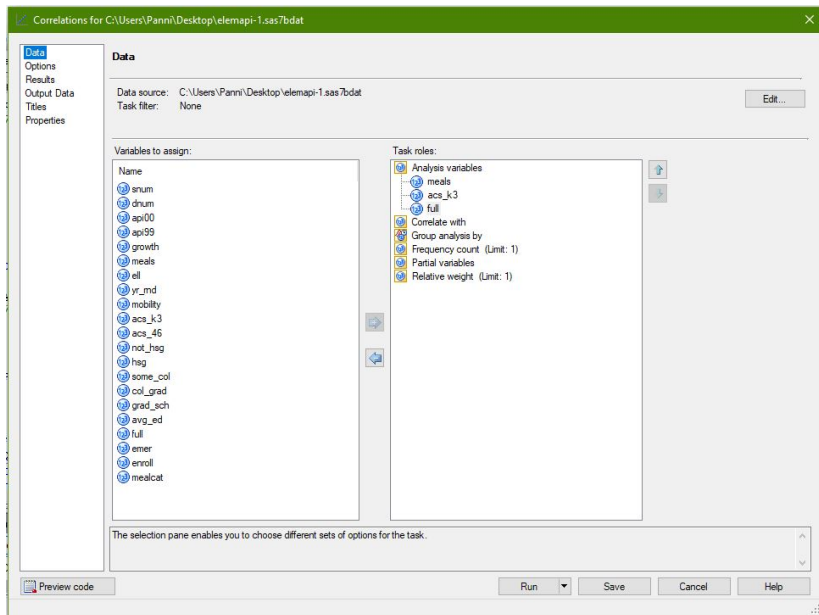
5. lépés – Az új adatok elemzése I.

- Az új adatokra ismét futtatunk egy Distribution Analysis-t, ahol láthatjuk, hogy reálisabb adatsort kaptunk.
- Futtatunk egy lineáris regressziót a jó adatokra, ahol láthatjuk, hogy nőtt az R^2 , azaz a nagyobb a regresszió varianciájának a megmagyarázott része → jobb a regresszió!
- Viszont a hipotézisvizsgálatnál az osztálylétszám még mindig nem szignifikáns magyarázó változó → hagyjuk el a regresszióból!
- Ekkor a modell magyarázó ereje nem csökkent jelentősen, viszont kevesebb adatot (változót) használtunk → így ez az eddigi legjobb regresszió!

5. lépés – Az új adatok elemzése II.

- Az *acs_k3* elhagyhatóságából következik, hogy ezt a változót megmagyarázza a másik kettő.
- Most vizsgáljuk meg a korrelációt és vegyük észre, hogy ez valóban így van.

Correlation



The screenshot shows the SAS 'Correlations' dialog box for the file 'C:\Users\Panni\Desktop\elemapi-1.sas7bdat'. The 'Data' tab is active, showing the data source and task filter. The 'Variables to assign' list includes variables like snum, dnun, api00, api99, growth, meals, ell, yr_md, mobility, acs_k3, acs_46, not_hsg, hsg, some_col, col_grad, grad_sch, avg_ed, full, emer, enroll, and mealcat. The 'Task roles' list includes 'Analysis variables' (with sub-roles meals, acs_k3, full), 'Correlate with', 'Group analysis by', 'Frequency count (Limit: 1)', 'Partial variables', and 'Relative weight (Limit: 1)'. A 'Preview code' button is at the bottom left, and 'Run', 'Save', 'Cancel', and 'Help' buttons are at the bottom right.

Correlations for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Data

Data source: C:\Users\Panni\Desktop\elemapi-1.sas7bdat
Task filter: None

Variables to assign:

- snum
- dnun
- api00
- api99
- growth
- meals
- ell
- yr_md
- mobility
- acs_k3
- acs_46
- not_hsg
- hsg
- some_col
- col_grad
- grad_sch
- avg_ed
- full
- emer
- enroll
- mealcat

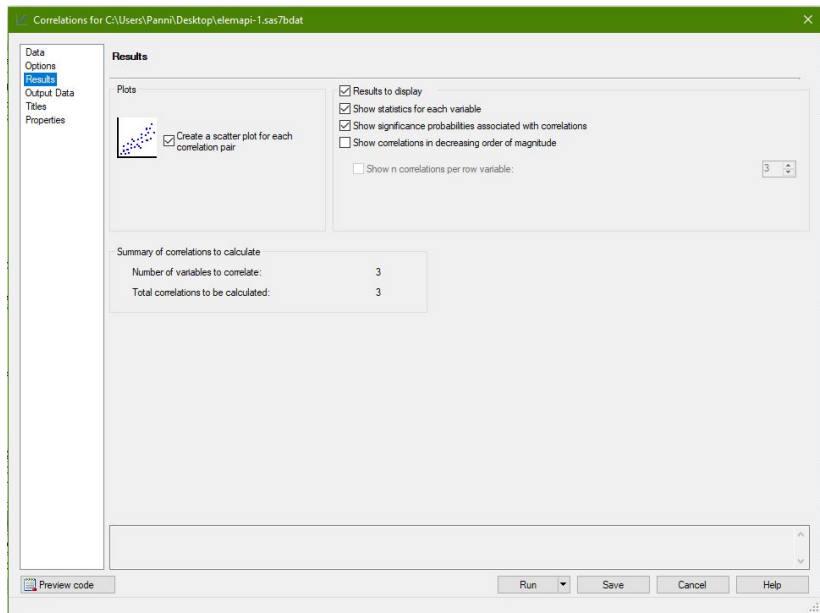
Task roles:

- Analysis variables
 - meals
 - acs_k3
 - full
- Correlate with
- Group analysis by
- Frequency count (Limit: 1)
- Partial variables
- Relative weight (Limit: 1)

The selection pane enables you to choose different sets of options for the task.

Preview code Run Save Cancel Help

Correlation



The screenshot shows the 'Correlations for C:\Users\Panni\Desktop\elemapi-1.sas7bdat' dialog box. The 'Results' tab is selected in the left-hand navigation pane. The 'Plots' section contains a scatter plot icon and a checked checkbox labeled 'Create a scatter plot for each correlation pair'. The 'Results to display' section has four checkboxes: 'Results to display' (checked), 'Show statistics for each variable' (checked), 'Show significance probabilities associated with correlations' (checked), and 'Show correlations in decreasing order of magnitude' (unchecked). Below these is a checkbox for 'Show n correlations per row variable:' with a spinner box set to '3'. A 'Summary of correlations to calculate' box shows 'Number of variables to correlate: 3' and 'Total correlations to be calculated: 3'. At the bottom, there are buttons for 'Preview code', 'Run', 'Save', 'Cancel', and 'Help'.

Correlations for C:\Users\Panni\Desktop\elemapi-1.sas7bdat

Data
Options
Results
Output Data
Titles
Properties

Results

Plots

Create a scatter plot for each correlation pair

Results to display
 Show statistics for each variable
 Show significance probabilities associated with correlations
 Show correlations in decreasing order of magnitude

Show n correlations per row variable: 3

Summary of correlations to calculate

Number of variables to correlate:	3
Total correlations to be calculated:	3

Preview code Run Save Cancel Help

Vége :)

Köszönjük a figyelmet!