

# Regresia segmentată și cauzalitatea

De la observație la inferență cauzală

Claudiu Papasteri

# Prezentarea

Urmăriți prezentarea pe:

<https://quarto.org/docs/presentations/>

sau scanând codul



Download:

[Prezentarea pdf](#)

[Codul din Github](#)

# Un pattern fără context (1)

► click pentru a vedea codul

Investigați codul:

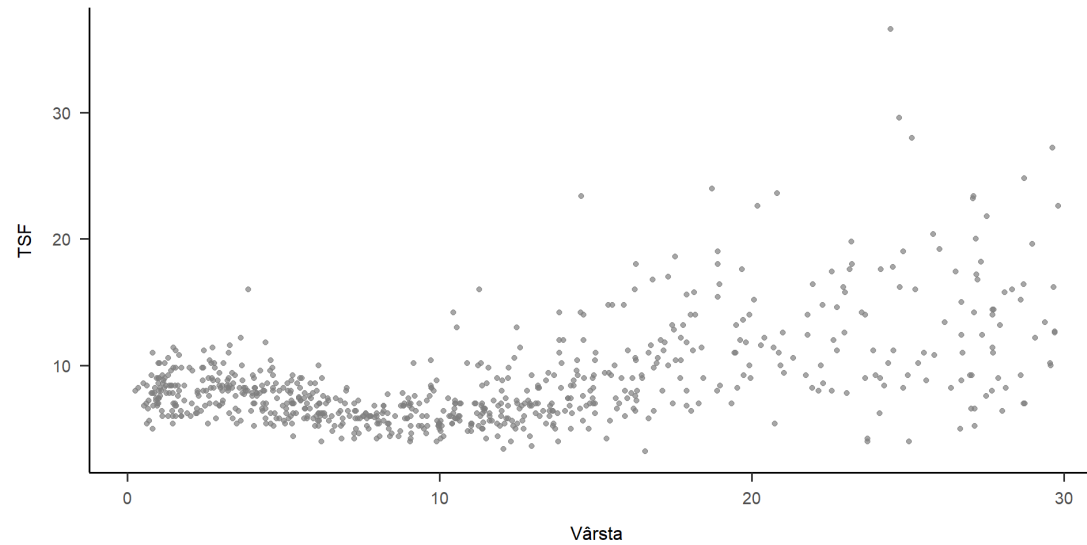
- Pachete & Setări
- Datele provin dintr-un pachet
- Se generează un grafic

# Un pattern fără context (2)

Ce vedeți?

Cum am putea modela?

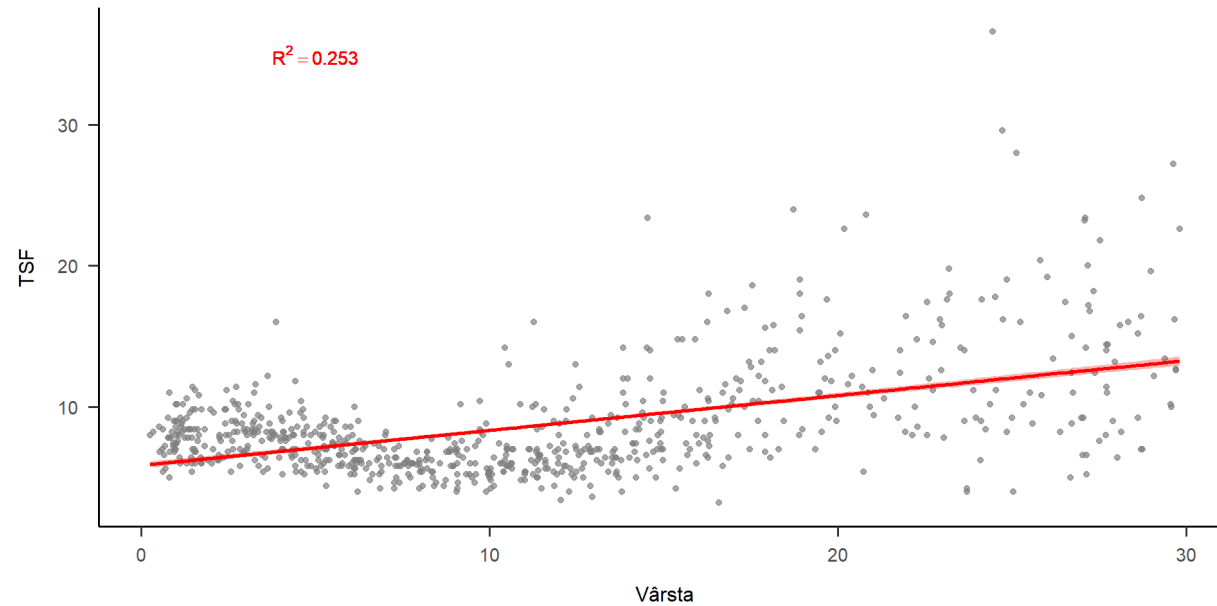
► click pentru a vedea codul



# Un pattern fără context (3)

Ce reprezintă linia? Este pattern-ul descris bine de o relație liniară?

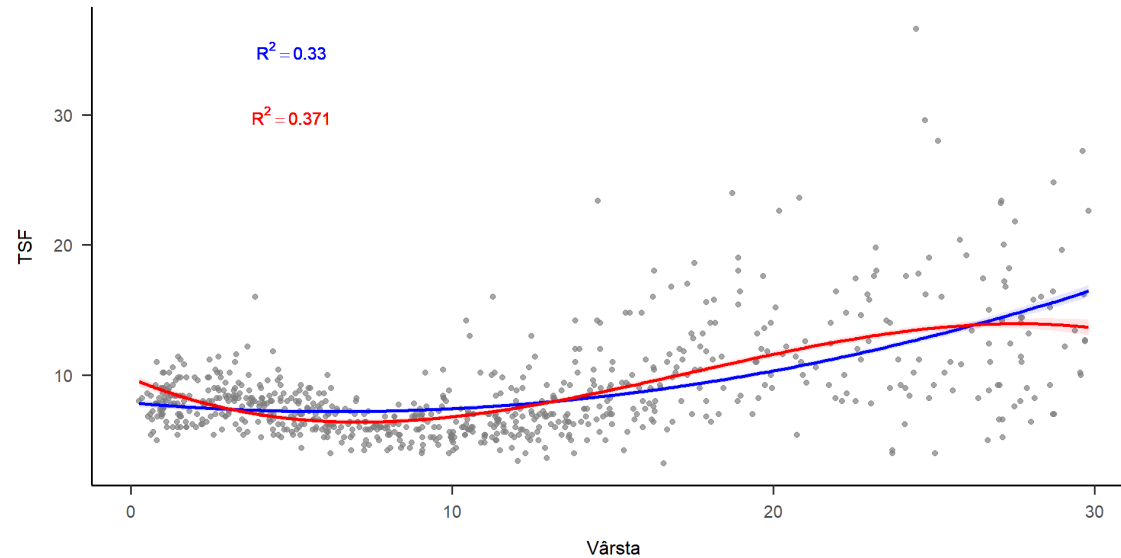
► click pentru a vedea codul



# Un pattern fără context (4)

Este patternul mai bine descris de o relație curbilinie (polinomială)? Care dintre trenduri, pătratic (**roșu**) sau cubic (**albastru**) se potrivește mai bine datelor?

► click pentru a vedea codul

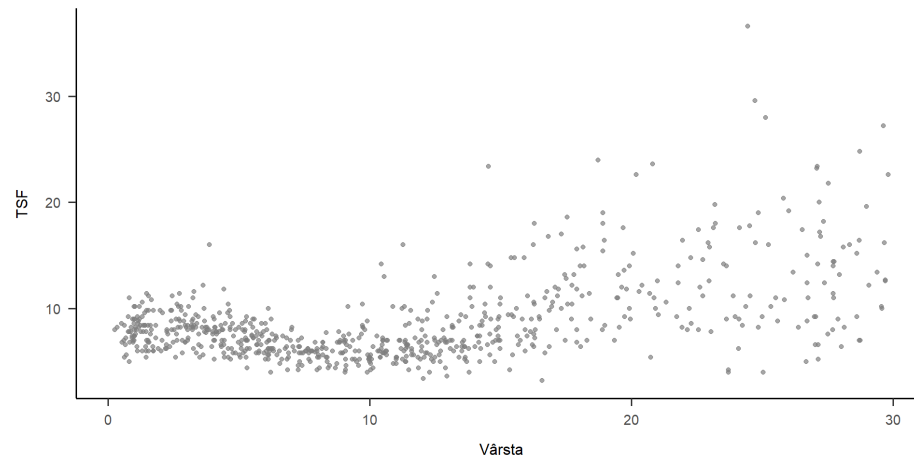


# Un pattern fără context (5)

Reinventăm regresia segmentată!

1. Împățim datele în grupe de vârstă (segmentăm variabila  $x$ ).
2. Potrivim regresii liniare pe fiecare set de date rezultat.

► click pentru a vedea codul



# Un pattern fără context (6)

Reinventăm regresia segmentată!

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \epsilon \quad \text{pentru } x \leq k_1 \\ y &= \beta_0 + \beta_1 x + \epsilon \quad \text{pentru } k_1 < x \leq k_2 \\ &\dots \\ y &= \beta_0 + \beta_1 x + \epsilon \quad \text{pentru } x \geq k_p \end{aligned}$$

Denumim  $k$  punctul de întrerupere (nod).

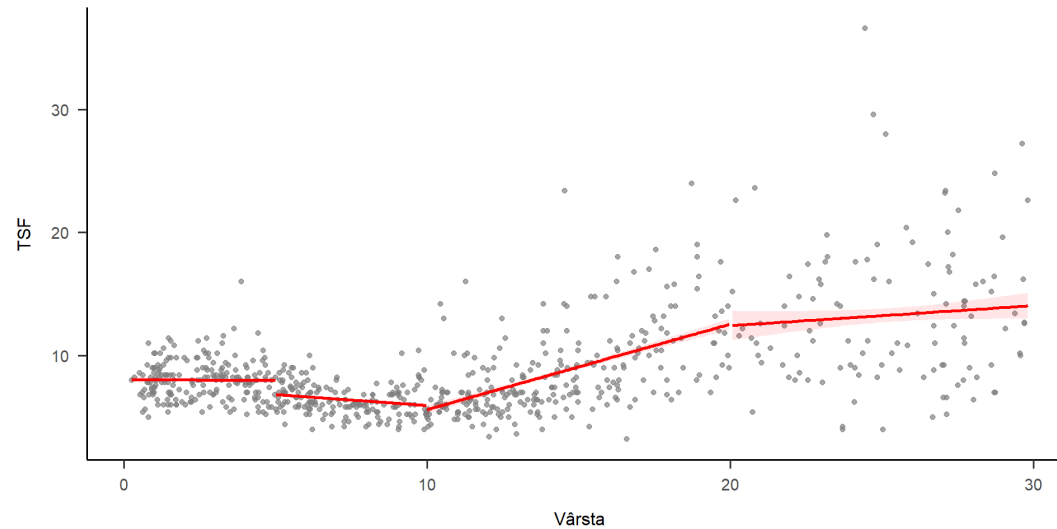
Rezultă  $p$  segmente.



# Un pattern fără context (7)

Reinventăm regresia segmentată, în cod!

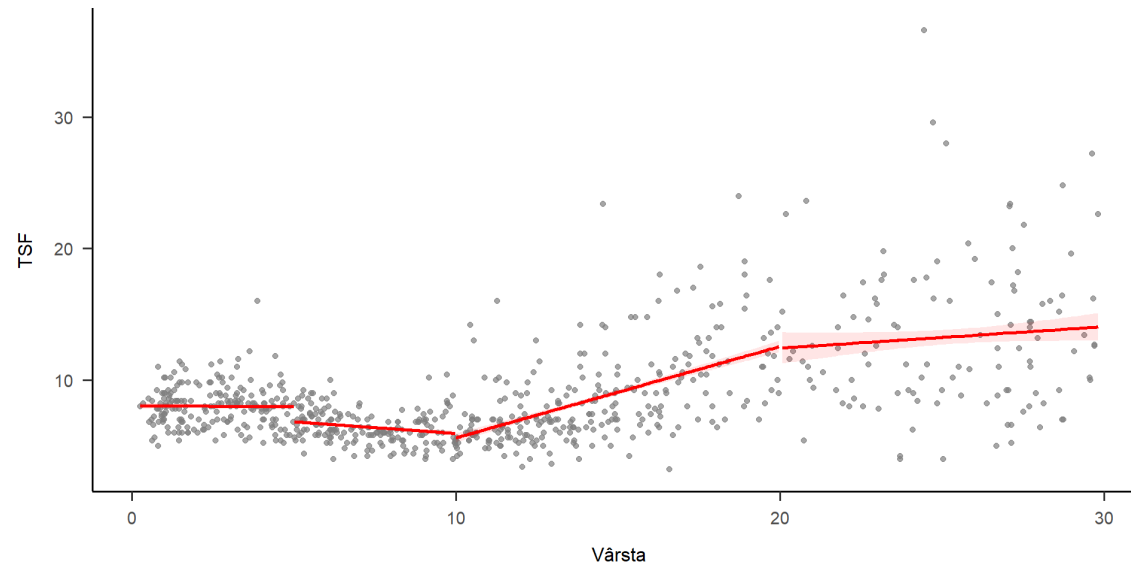
► [click pentru a vedea codul](#)



# Un pattern fără context (8)

Reinventăm regresia segmentată, în cod!

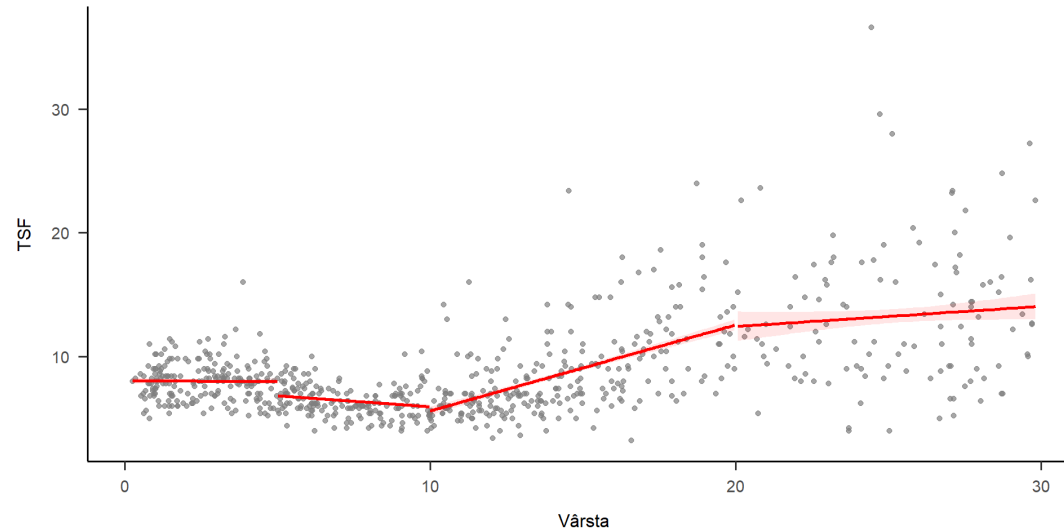
Ce observați despre segmentele din Regresia Segmentată?



# Un pattern fără context (9)

De fapt, ceea ce este denumit în literatură Regresie Segmentată nu este Regresie Segmentată, ci Regresie cu Spline-uri Liniare.

Spline-urile permit o interpolare netedă și continuă între punctele de întrerupere (noduri).  
Între noduri este calculată o regresie (*polinomială*).



# Un pattern fără context (10)

Re-reinventăm regresia segmentată!

1. Împățim datele în grupe de vârstă (segmentăm variabila  $x$ )  $\Rightarrow k$  noduri.
2. Folosim o funcție treaptă încât să obținem o singură ecuație de regresie care păstrează doar un intercept  $\beta_0$

$$y = \beta_0 + \beta_1 x + \beta_2(x - k_1) + \beta_3(x - k_2) + \dots + \beta_{p-1}(x - k_{p-1})$$

$$Unde \quad (x - k)_+ = \begin{cases} 0 & \text{dacă } x < k \\ x - k & \text{dacă } x \geq k \end{cases}$$

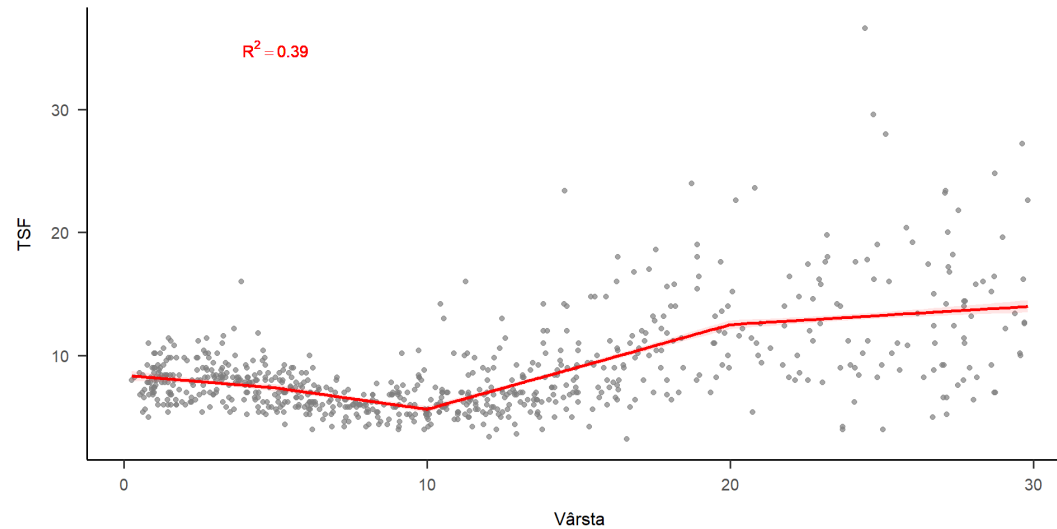
Denumim  $k$  punctul de întrerupere (nod).

Rezultă  $p$  segmente.

# Un pattern fără context (11)

Re-reinventăm regresia segmentată, în cod!

► click pentru a vedea codul

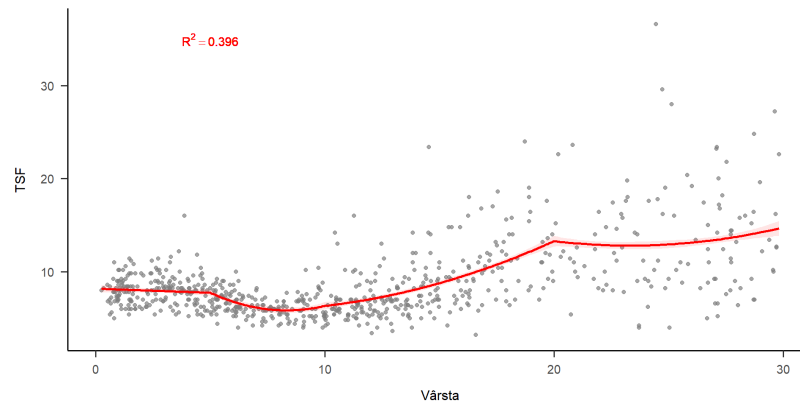


# Un pattern fără context (12)

**Bonus:** Mai sus scria că regresiiile între noduri sunt *polinomiale*, iar polinomul, poate desigur să nu fie de ordinul I.

Optimizând numărul  $k$  de noduri și ordinul polinomului obținem o metodă foarte robustă, deși nesofisticată, de învățare a pattern-urilor din date. Un exemplu de Regresie cu Spline-uri Polinomiale de ordin II:

► [click pentru a vedea codul](#)



# Contextul datelor, elucidarea misterului

Măsurăm constructe:

- **TSF** = grosimea pliului cutanat al tricepsului, o măsurătoare economică și convenabilă pentru a evalua obezitatea
  - Validitate: avantaj față de alți indici are avantajul de a reprezenta distribuția grăsimii.
  - Utilitate predictivă: la fiecare creștere cu 1 mm a grosimii TSF, riscul de deces scade cu 4%, riscul de deces

**Vârsta** codează timpul în unități discrete dacă am presupune că indivizii sunt interșanjabili. Totuși multe variab

# Date longitudinale în Psihologie

Deoarece cea mai mare parte a psihologiei implică așa-numite procese non-ergodice, cercetarea pe eșantion mari nu poate oferi informații fiabile despre procesele la nivel individual. ([Hamaker, 2012](#)).

- O descriere simplificată a non-ergotismului: individul, în timp, nu obține rezultatul mediu al grupului.

Monitorizare în timp real al sistemelor idiografice

- Pentru o discuție a utilizării conceptului și complianței cu metodologia în context psihoterapeutic vedeți ([Schiepek et al., 2016](#)).



# Datele longitudinale

Datele disponibile de la [Journal of Open Psychology Data](#): ([Heino, 2022](#))

- Monitorizare dinamică cognitivă (test Stroop) după mindfulness
- Un singur individ peste 900 de zile, meditație zilnică de 20 minute

[Repozitoriu OSF](#)

Datele se găsesc și în Github-ul prezentării.

# Lupta cu date longitudinale (1)

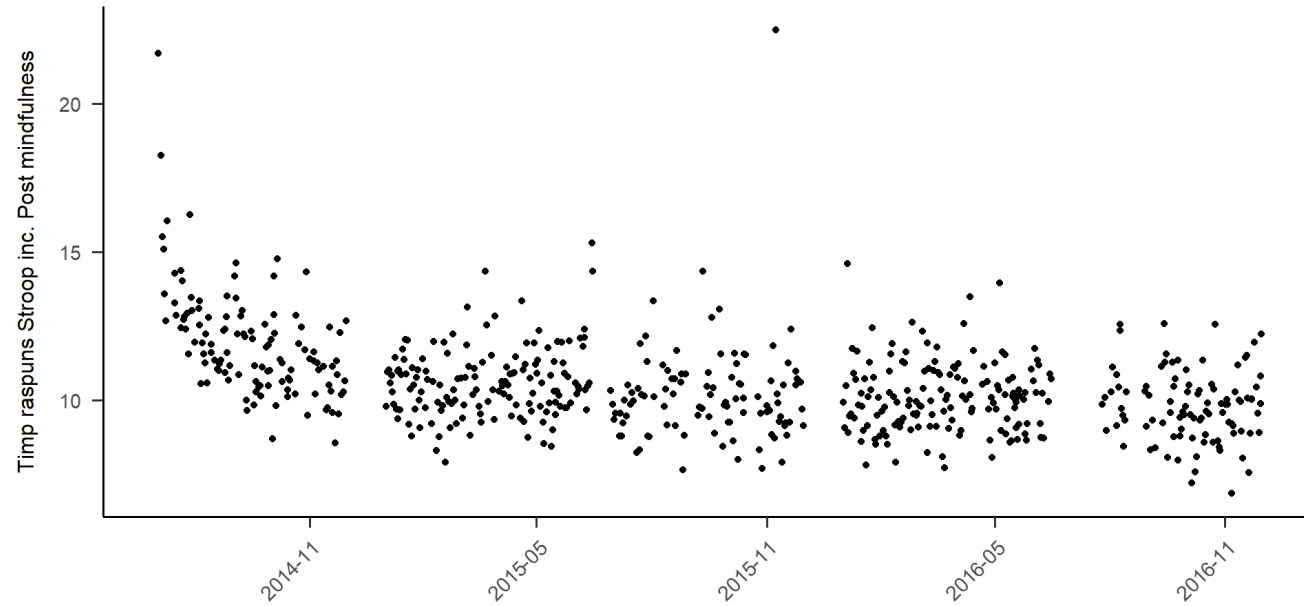
► click pentru a vedea codul

Investigați codul:

- Pachete & Setări
- Datele sunt citite din repositoryul OSF
- Datele sunt curățate și transformate
- Se generează un grafic

# Lupta cu date longitudinale (2)

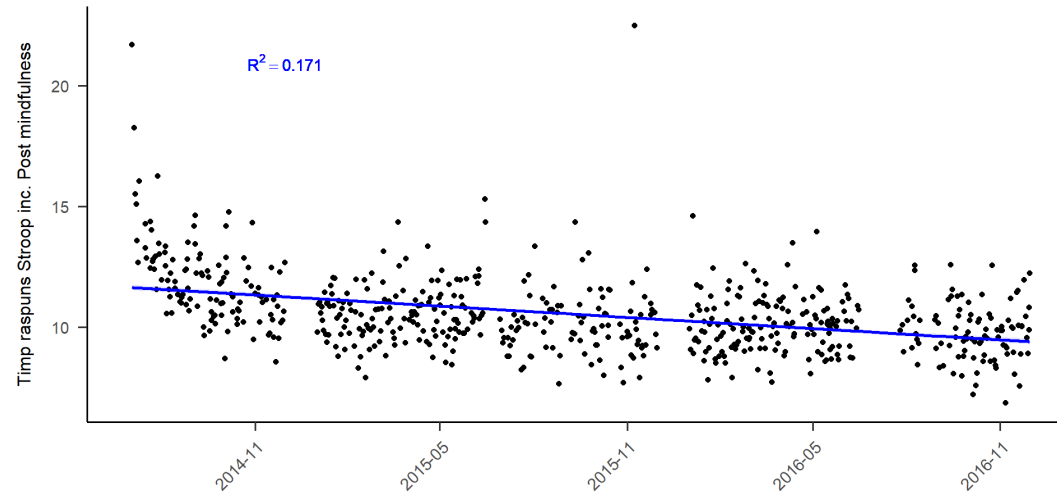
Ce observați?



# Regresia Liniară

Putem modela mai bine?

► click pentru a vedea codul



# Regresia segmentată

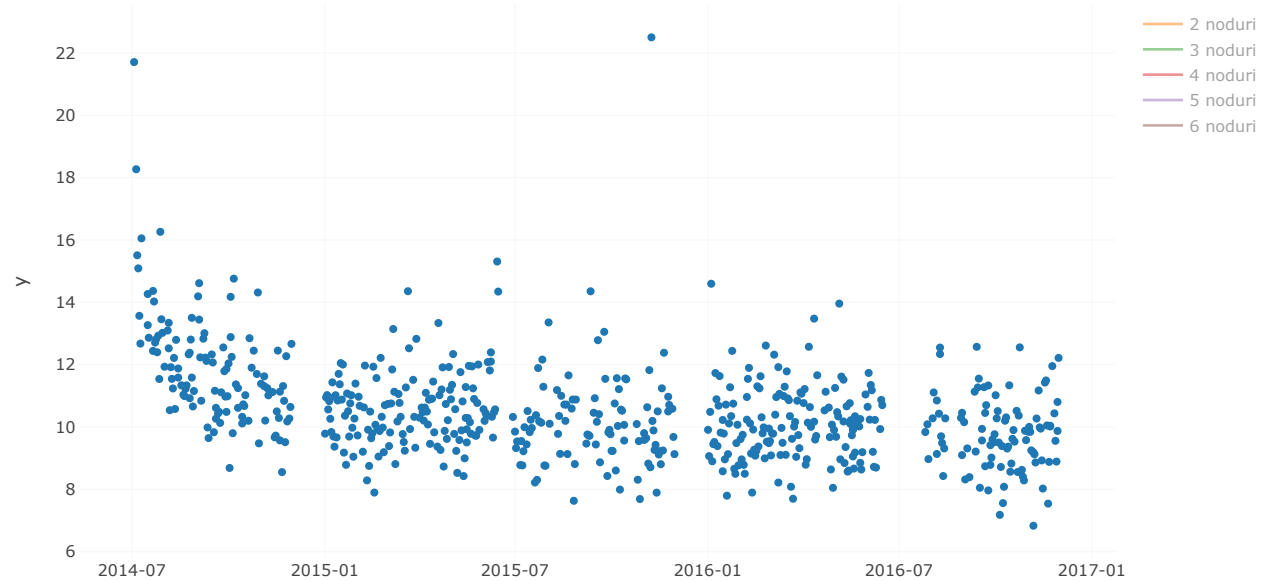
De această dată nu o vom mai coda “manual” (folosind doar *base R*), ci vom folosi pachetul [lspline](#) pentru a obține noduri în urma împărțirii datelor în intervale egale de timp.

► click pentru a vedea codul

# Interacționați cu datele

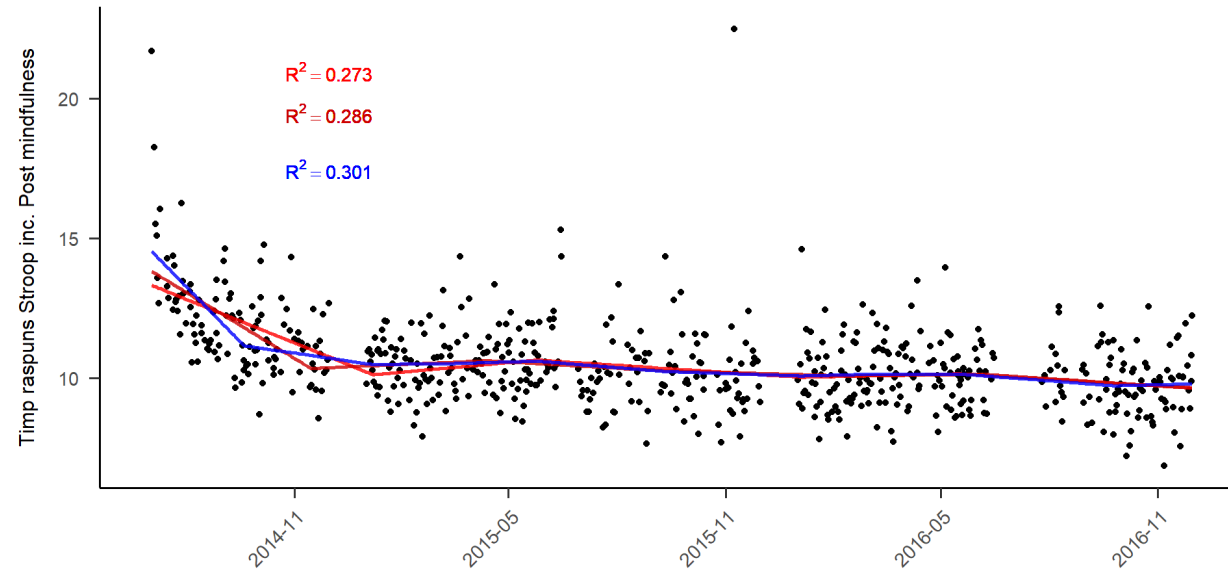
Câte noduri să folosim?

► click pentru a vedea codul



# Câte noduri?

► click pentru a vedea codul



# Pe ce bazăm inferența

Inferență bazată pe model

V.S.

Inferență bazată pe design



# Cauzalitatea (după Hume)

$$C \rightarrow E$$

Cauza și efectul sunt interconectate

$$C \rightarrow E$$

Cauza precede (temporal) efectul

$$\begin{array}{c} C \rightarrow E \\ C \rightarrow E \\ C \rightarrow E \end{array}$$

Cauza și efectul covariază consistent

$$\begin{array}{c} A \\ C \rightarrow E \end{array}$$

Nu există alte explicații alternative

# Regresia segmentată și cauzalitatea (1)

## Design-ul de Discontinuitate în Regresie (RDD)

- cvasi-experimental pre-post
- manipularea variabilei independente poate apărea natural în mediu => inferență cauzală din date observaționale
- NU vom discuta despre ea la acest curs, dar puteți [citiți mai multe](#)

# Regresia segmentată și cauzalitatea (2)

## Studiile Experimentale cu un Singur Caz (SCED)

- manipularea variabilei independente
- măsurători repetate
- calitatea dovezilor cauzale echivalentă cu cea din studii clinice randomizate (RCT)

# Regresia segmentată în SCED

0.1177/002246698501900404 • Corpus ID: 145752030

# Methodology for the Quantitative Synthesis of tra-Subject Design Research

ter, R. Skiba, A. Casey • Published 1 December 1985 • Psychology • Journal of Special Education

tigators using quantitative synthesis methodology have as yet been unable to include the results from single-experiments in their analyses, in large part because of the lack of a suitable statistical methodology. The opment of a regression model that can be used to generate effect sizes for both changes in slope and ges in level occurring as a result of the treatment intervention is outlined. Analysis of lag one autocorrelation g the residuals suggests that serial correlation is not a serious problem. The regression approach provides a actory fit of the data, and may thus provide a basis for the generation of effect sizes from single-case imental data. [Collapse](#)

(Center et al., 1985)

# Istoria modelului pentru SCED

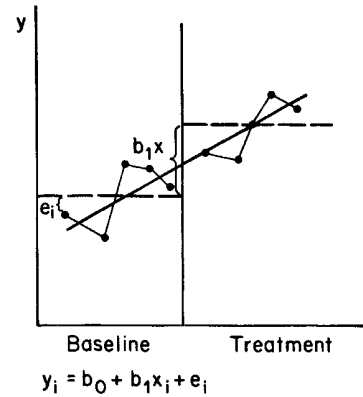


Figure 1. Parameters of the simple ANOVA model. Dashed line represents the mean for each phase; solid line, the overall slope of the data.

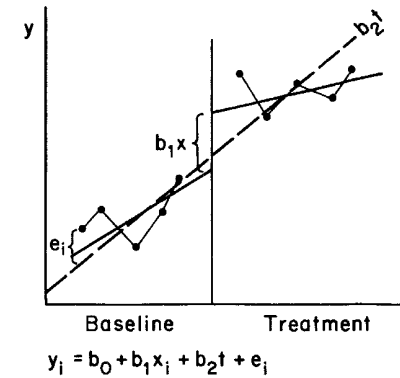


Figure 2. Parameters of the [overall slope] model presented by Gorsuch (1983). Dashed line represents the regression slope for the entire data set; solid line, the within-phase slopes.

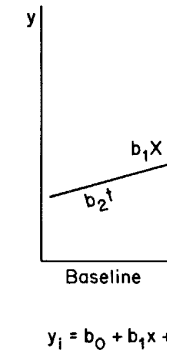


Figure 3a. Parallel line regression model. 3b. Not only

Model ANOVA

Model cu o pantă globală

Model regresie segmentată

# Istoria modelului în cod (1)

Codați cu mine pașii:

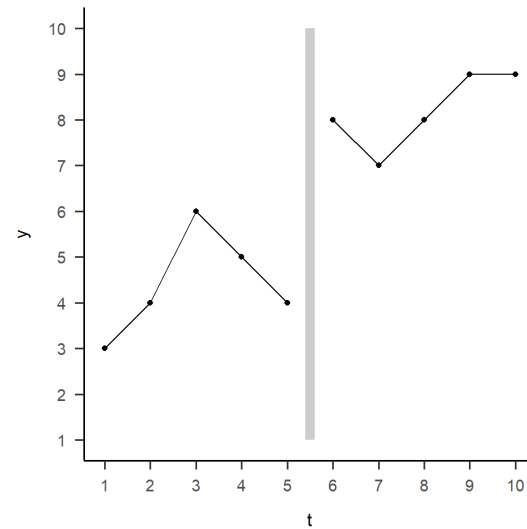
- Pachete & Setări
- Date
- Calcul statistici pentru grafice

► click pentru a vedea codul

# Istoria modelului în cod (2)

Plotarea datelor, un schelet pentru restul graficelor.

► click pentru a vedea codul

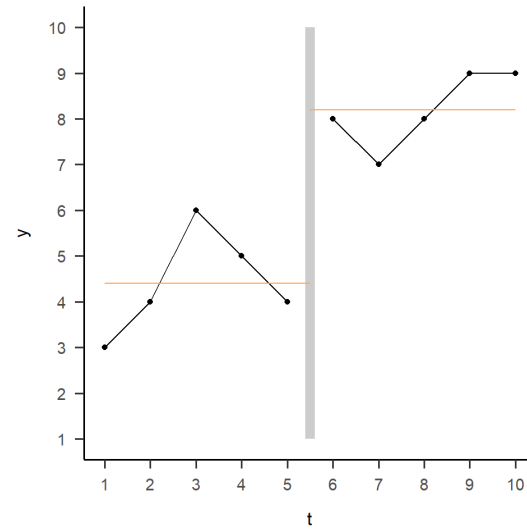




# Istoria modelului în cod (3)

Mediile celor două faze.

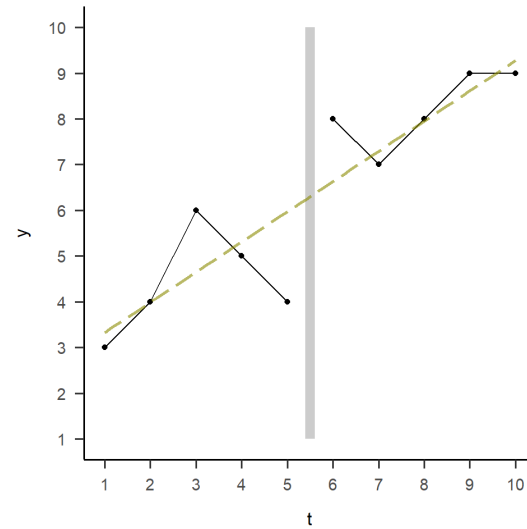
► [click pentru a vedea codul](#)



# Istoria modelului în cod (4)

Panta globală (peste datele din ambele faze).

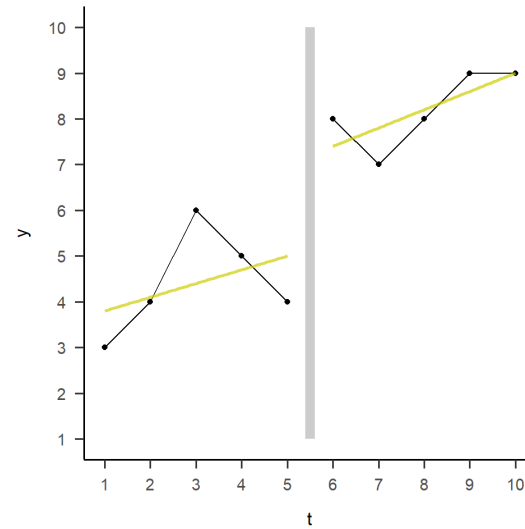
► click pentru a vedea codul



# Istoria modelului în cod (5)

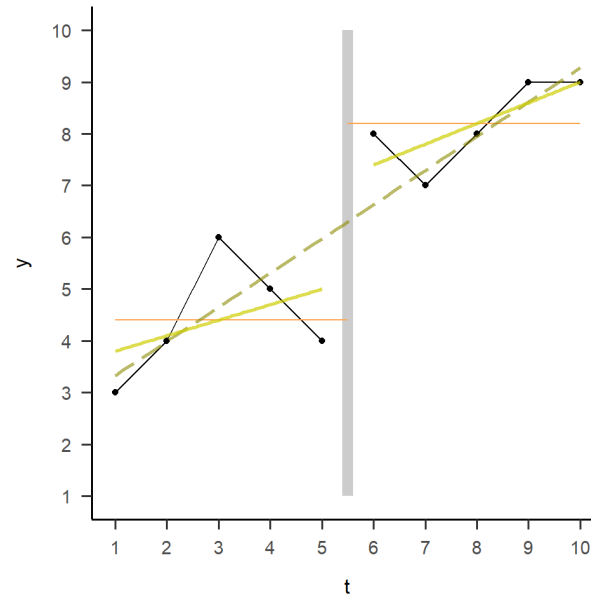
Pantele individuale ale celor două faze.

► click pentru a vedea codul



# Istoria modelului în cod (6)

Ce arată graficul?

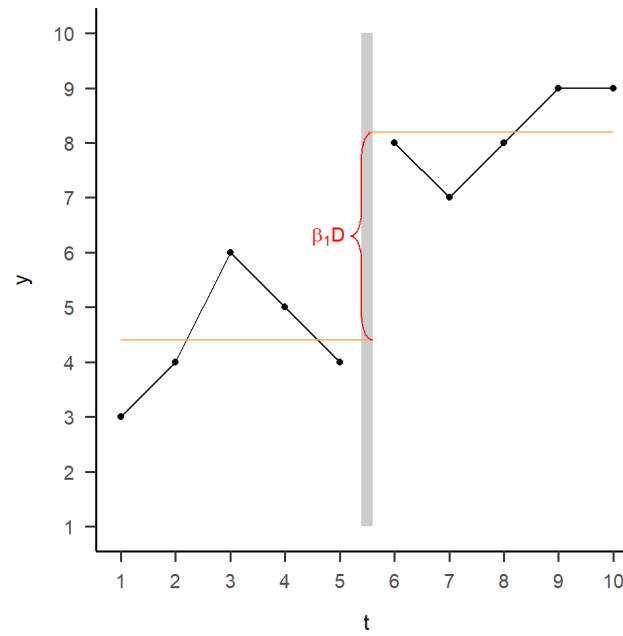


# Istoria modelului în cod (7)

Cum am obținut graficul?

```
1 # Plot all together
2 plot_skeleton +
3   geom_segment(aes(x = 1, xend = 5.5, y = mean_A, yend = mean_A),
4                 color = "tan1") +
5   geom_segment(aes(x = 5.5, xend = 10, y = mean_B, yend = mean_B),
6                 color = "tan1") +
7   stat_smooth(data = df, aes(x = t, y = y), geom = "line",
8                formula = y ~ x, method = "lm", se = FALSE,
9                linetype = "longdash", size = 1, alpha = 0.6,
10               color = "yellow4") +
11   stat_smooth(data = df_A, aes(x = t, y = y), geom = "line",
12                formula = y ~ x, method = "lm", se = FALSE,
13                alpha = 0.7, size = 1, color = "yellow3") +
14   stat_smooth(data = df_B, aes(x = t, y = y), geom = "line",
15                formula = y ~ x, method = "lm", se = FALSE,
16                alpha = 0.7, size = 1, color = "yellow3")
```

# Modelul ANOVA

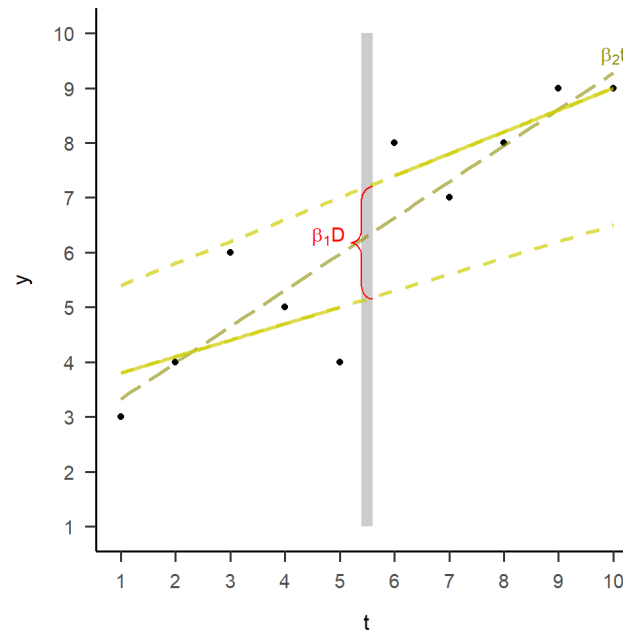


$$y = \beta_0 + \beta_1 D + \epsilon$$

# Modelul ANOVA

- diferență în medii (nivel) =  $\beta_1$

# Modelul cu o Pantă Globală



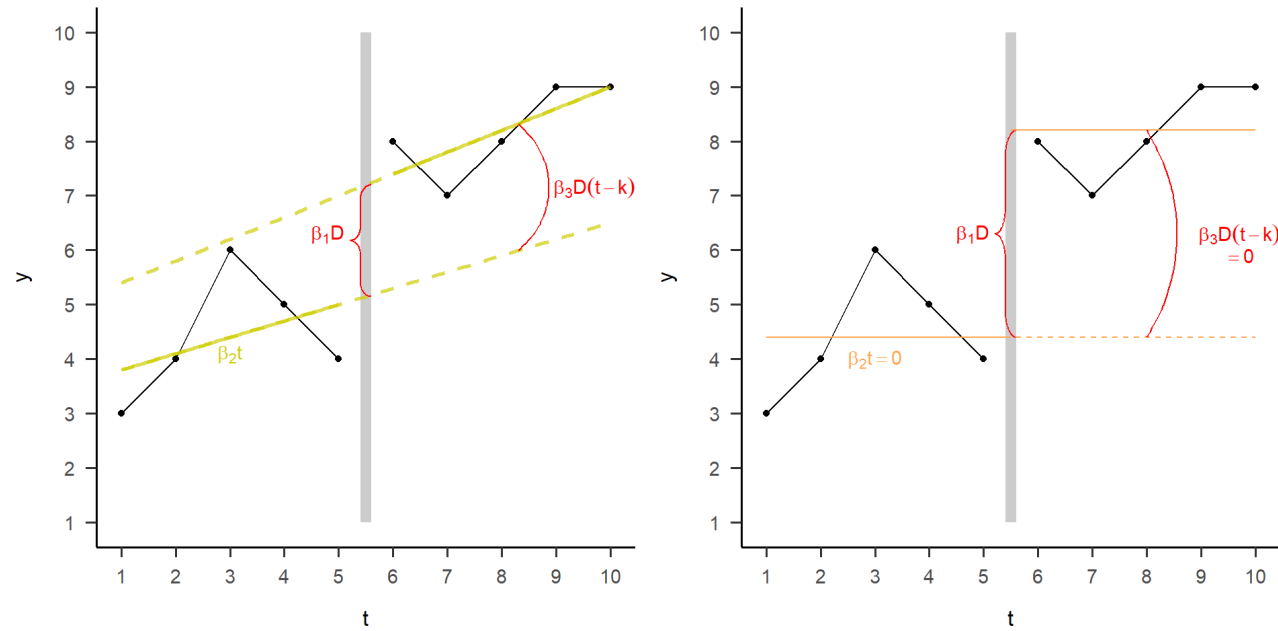
$$y = \beta_0 + \beta_1 D + \beta_2 t + \epsilon$$



# Modelul cu o Pantă Globală

- ([Gorsuch, 1983](#))
- diferență în medii (nivel) =  $\beta_1$ , dar doar dacă nu există trend (adică  $\beta_2 = 0$ )
- ia în calcul trendul, dar presupune trend identic între faze

# Modelul de Regresie Segmentată



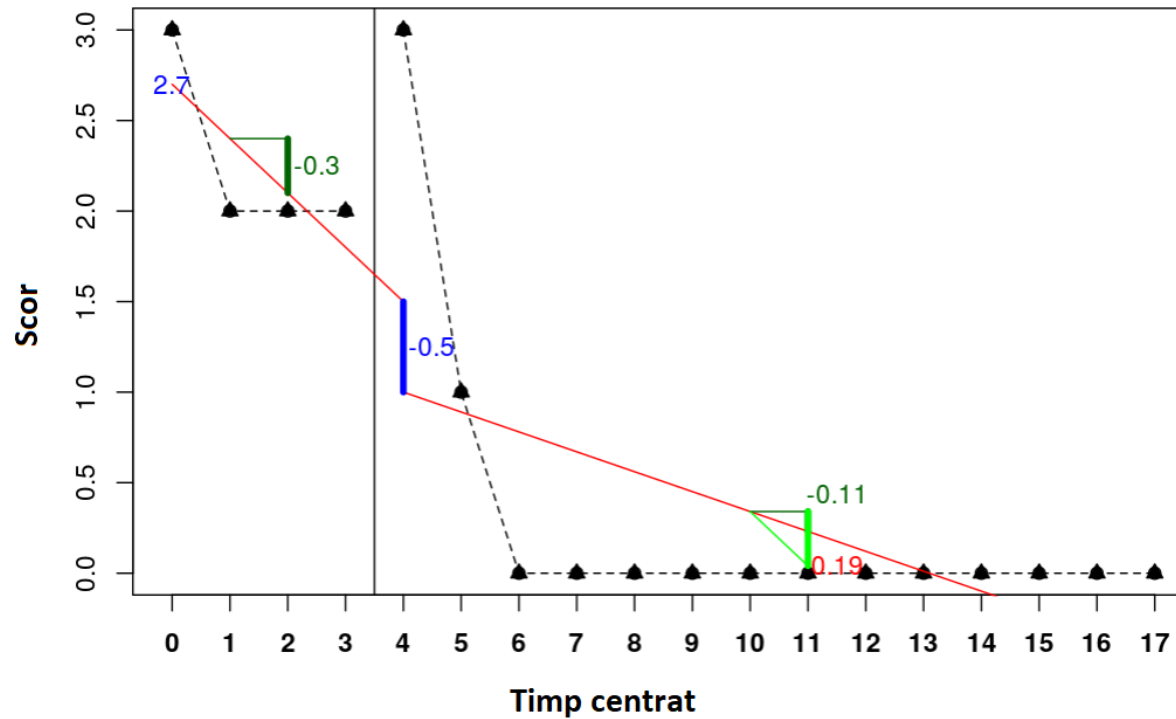
$$y = \beta_0 + \beta_1 D + \beta_2 t + \beta_3 D(t - k) + \epsilon$$

# Modelul de Regresie Segmentată

- ([Center et al., 1985](#))
- diferență în medii (nivel) =  $\beta_1 D$
- trend baseline =  $\beta_2 t$
- termen de interacțiune ce cuantifică diferența în pantă între cele două faze =  $\beta_3 D(t - k)$
- unde  $k$  = nr observații în faza A (baseline)

# Interpretare

## Efecte nestandardizate



$$y = \beta_0 + \beta_1 D$$

$\beta_0$  este interceptul (scorul evaluat la  $t = 0$  și  $D = 0$ ), deci indică valoarea de început a liniei de regresie în

$\beta_1$  poate fi interpretat ca schimbarea nivelului dintre faza A și faza B necounfounduit cu posibilele efecte c

$\beta_2$  captează trendul din faza A.

$\beta_3$ , care este de obicei parametrul cel mai de interes, reprezintă modificarea în trend de la faza A la faza B.



# Datele SCED

Datele provin din celebrul studiu realizat de ([Singh et al., 2007](#)), digitalizate de Rumen Manolov.

## Variable:

tier	Id numeric participant
id	Nume participant
time	Index al momentului măsurătorii ( $t_0 = 1$ ).
phase	Variabilă dummy: 0 tratamentul nu a început (baseline), 1 tratamentul a început.
score_physical	Scorul participantului la agresivitate fizică.
score_verbal	Scorul participantului la agresivitate verbală.

Download

[Repozitoriu OSF](#)

Datele participantului “Michael” utilizate în această prezentare se găsesc și în Github-ul prezentării.

# Preambul în cod

► click pentru a vedea codul

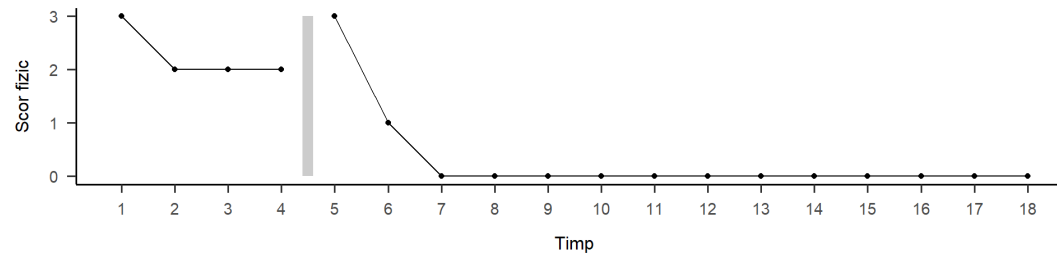
Investigați codul:

- Pachete & Setări
- Datele provin dintr-un pachet
- Se generează un grafic

# Vizualizarea datelor

Ce observați?

► click pentru a vedea codul



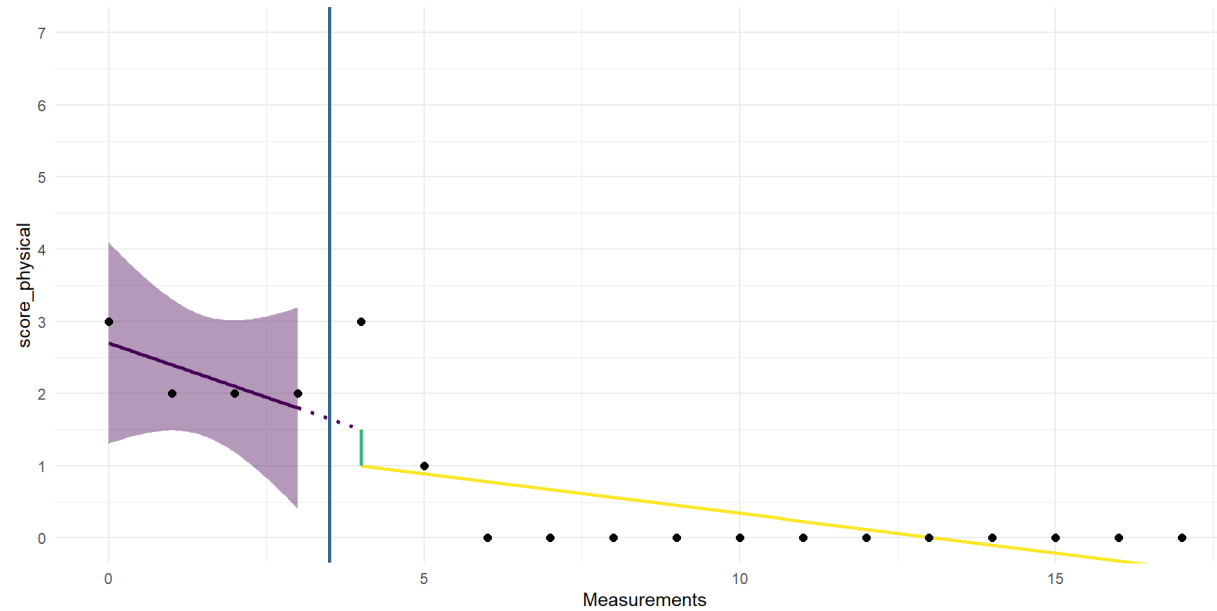


# Analize cu pachetul **scda**

► click pentru a vedea codul

# Grafic cu pachetul **scda**

► click pentru a vedea codul



# Rezultate cu pachetul **scda**

► click pentru a vedea codul

Piecewise Regression Analysis (N = 18)

Model statistics:

Model deviance:	6.41
R squared for null model:	0.63
R squared for test model:	0.7
R squared based effect size:	0.19
Effect size (delta_t):	3.36
Standardized effect size (delta_ts):	5.47
Effect size (delta):	0.74
Standardized effect size (delta_s):	1.2

Effect size evaluated at point:	18
---------------------------------	----

Regression coefficients:

# Mărime a efectului cu pachetul **scda**

$\delta_t$  a lui (Swaminathan et al., 2014) compară scorurile prezise la un anumit punct de măsurare (t) din faza B cu predicția din faza A extrapolată la t.

► click pentru a vedea codul

```
(Intercept)  
3.361538
```

# Regresie segmentată generalizată cu pachetul **scda**

Extindere a modelului utilizată în designuri mai complexe (ex. ABA cu retragere, ABAB cu inversiune).

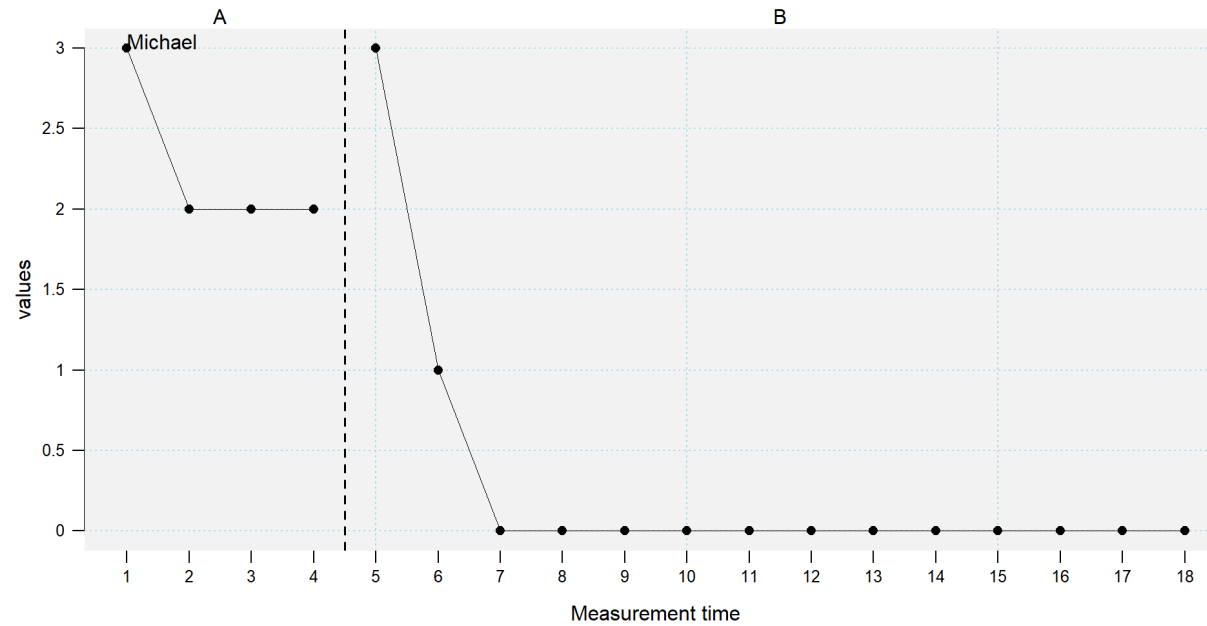
► [click pentru a vedea codul](#)

# Analize cu pachetul **scan**

► click pentru a vedea codul

# Grafic cu pachetul **scan**

► click pentru a vedea codul



# Rezultate cu pachetul **scan**

► click pentru a vedea codul

Piecewise Regression Analysis

Dummy model: H-M

Fitted a gaussian distribution.

$F(3, 14) = 11.07$ ;  $p = 0.001$ ;  $R^2 = 0.703$ ; Adjusted  $R^2 = 0.640$

	B	2.5%	97.5%	SE	t	p	delta	$R^2$
Intercept	3.00	1.376	4.624	0.829	3.620	0.003		
Trend mt	-0.30	-0.893	0.293	0.303	-0.991	0.338	0.0208	
Level phase B	-0.50	-2.258	1.258	0.897	-0.557	0.586	0.0066	
Slope phase B	0.19	-0.409	0.790	0.306	0.621	0.544	0.0082	

Autocorrelations of the residuals

lag	cr
1	0.22



# Autocorelație și covariate cu pachetul **scan**

► click pentru a vedea codul

```
Multivariate piecewise linear model
```

```
Dummy model: H-M
```

```
Coefficients:
```

	values	covariate
(Intercept)	3.00	6.500
Trend	-0.30	-1.000
Level Phase B	-0.50	0.271
Slope Phase B	0.19	0.870

```
Formula: y ~ 1 + mt + phaseB + interB
```

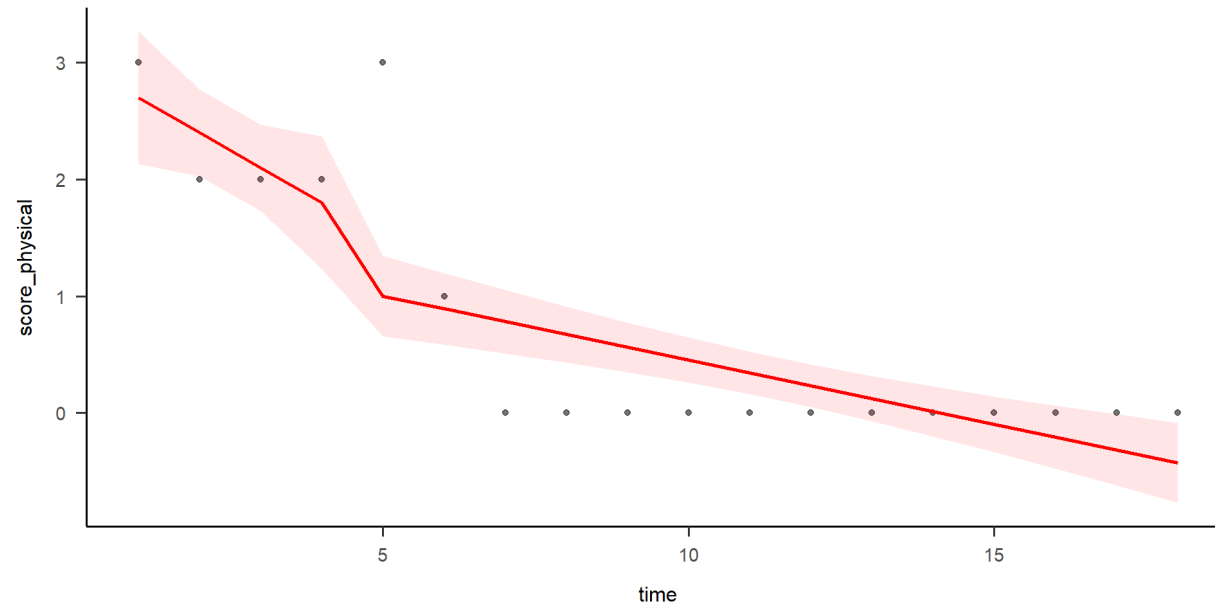
```
Type III MANOVA Tests: Pillai test statistic
```

	Df test	stat	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.558	8.22	2	12	0.0040 **

# Reproducem fără pachete

Totul în base R.

► [click pentru a vedea codul](#)



# Reproducem fără pachete

Observăm o diferență în estimare interceptului între `scda` și `scan` datorată centrării timpului. Puteți citi mai multe despre recomandări de codificare a timpului în SCED în (Huitema & Mckean, 2000) (mai ales schem standard H-M ce poartă numele autorilor).

În base R putem reproduce ambele rezultate.

La fel ca `scda`

```
Estimate Std. Error    t value
(Intercept)
2.7000000  0.5661223  4.7692871
phase
-0.5000000  0.8969429 -0.5574491
time
-0.3000000  0.3026051 -0.9913910
I(phase * (time - k_time) * (time >= k_time))
0.1901099  0.3059124  0.6214521
```

```
Pr(>|t|)
(Intercept)
0.0002993711
phase
0.5860238465
time
0.3383210448
I(phase * (time - k_time) * (time >= k_time))
0.5442892970
```

La fel ca `scan`

```
Estimate Std. Error    t value
(Intercept)
3.0000000  0.8287183  3.6200482
phase
-0.5000000  0.8969429 -0.5574491
time
-0.3000000  0.3026051 -0.9913910
I(phase * (time - k_time) * (time >= k_time))
0.1901099  0.3059124  0.6214521
```

```
Pr(>|t|)
(Intercept)
0.002785467
phase
0.586023846
time
0.338321045
I(phase * (time - k_time) * (time >= k_time))
0.544289297
```

# Analize în GUI

<https://manolov.shinyapps.io/Regression/>

### Regression analysis options

## User input

Use a .txt file with 'score' and 'phase' as column names

**Load data file**

Upload complete

**Separator**

☐ Comma

☒ Tab

☐ Space

**Minimal possible value**

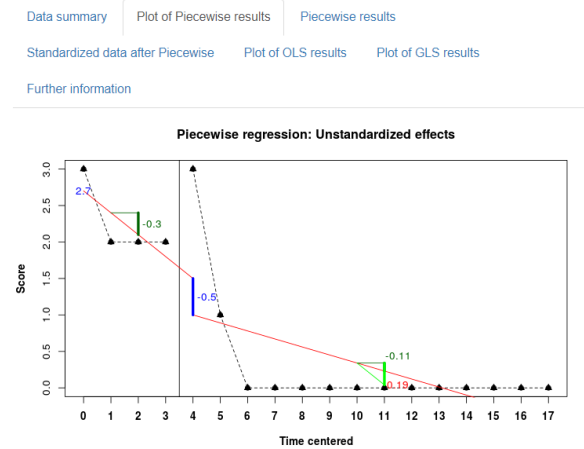
**Maximal possible value**

**For GLS, specify whether to transform data directly or only if autocorrelation is statistically significant**

☒ ifsig

☐ directly

Visited: 103



# Bibliografie

- Center, B. A., Skiba, R. J., & Casey, A. (1985). A Methodology for the Quantitative Synthesis of Intra-Subject Design Research. *The Journal of Special Education*, 19(4), 387–400.  
<https://doi.org/10.1177/002246698501900404>
- Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (n of 1) data. *Behavioral Assessment*, 5, 141–154.
- Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In *Handbook of research methods for studying daily life*. (pp. 43–61). The Guilford Press.
- Heino, M. T. J. (2022). Cognitive Dynamics of a Single Subject: 1428 Stroop Tests and Other Measures in a Mindfulness Meditation Context Over 2.5 Years. *Journal of Open Psychology Data*, 10. <https://doi.org/10.5334/jopd.51>
- Huitema, B. E., & Mckean, J. W. (2000). Design Specification Issues in Time-Series Intervention Models. *Educational and Psychological Measurement*, 60(1), 38–58.  
<https://doi.org/10.1177/00131640021970358>
- Li, W., Yin, H., Chen, Y., Liu, Q., Wang, Y., Qiu, D., Ma, H., & Geng, Q. (2022). Associations between adult triceps skinfold thickness and all-cause, cardiovascular and cerebrovascular mortality in NHANES 19992010: A retrospective national study. *Frontiers in Cardiovascular Medicine*, 9. <https://doi.org/10.3389/fcvm.2022.858994>
- Schiepek, G., Aichhorn, W., Gruber, M., Strunk, G., Bachler, E., & Aas, B. (2016). Real-time monitoring of psychotherapeutic processes: Concept and compliance. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00604>
- Singh, N. N., Lancioni, G. E., Winton, A. S. W., Adkins, A. D., Wahler, R. G., Sabaawi, M., & Singh, J. (2007). Individuals with Mental Illness Can Control their Aggressive Behavior Through Mindfulness Training. *Behavior Modification*, 31(3), 313–328.  
<https://doi.org/10.1177/0145445506293585>
- Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, 52(2), 213–230.



