Globalization and Higher Education in Economics and Business Administration
Iasi, GEBA 18-20 October 2012

# BRINGING NEW OPPORTUNITIES TO DEVELOP STATISTICAL SOFTWARE AND DATA ANALYSIS TOOLS IN ROMANIA

**Nicoleta CARAGEA**
Senior Expert, National Institute of Statistics, Romania/Lecturer at Ecological University of Bucharest
**Ciprian Antoniade ALEXANDRU**
Senior Lecturer at Ecological University of Bucharest
**Ana-Maria DOBRE**
Expert, National Institute of Statistics

# R - a new challenge at academic level

Why R?

- Data analysis tool - high inertia to change:

- level of knowledge that young people receive in the academic level

- tradition of teachers in keeping poor updated topics

- Re-shape the thinking at university level

- present to students the opportunity to use the state-of-the-art programming technology for data analysis

# R - overview

- Data analysis tool developed by statisticians for statisticians
- Statistical programming language created by two academicians in 1993 in New Zealand and released in 1996
- Free and open-source software
- User-friendly Graphical User Interfaces: RStudio, Deducer, Revolution Analytics, Red-R, JGR (Java GUI for R), SciViews-R
- In about three years the R's users will exceed the number of users of SAS and SPSS

RStudio

File Edit Code View Project Workspace Plots Tools Help

Go to file/function

Project: (None)

calea scurta 2007.R | calea scurta 2008.R | calea scurta 2009.R | calea scurta 2011.R* | amigo1_

Source on Save | Run | Source

```
33  amigo1 <- aggregate(amigo[,c( "judet","absent","coef","sex_1",   "sex_2",   "med
34
35  amigo2 <- aggregate(amigo[,c( "judet","absent","coef","sex_1",   "sex_2",   "med
36
37
38  head(rpl_judete)
39  head(amigo1)
40
41
42
43  #estimam migratia pe judete
44
45
46  summary(fit.lme <- lme(absent~edu_inf+mediu_r+gr_5, data = amigo1, random = ~1
47
48
49
50  d.data <- rpl_judete
51  head(d.data)
52  result <- eblup.mse.f.wrap(domain.data = d.data, lme.obj = fit.lme)
53  result
54
55  estimatori<- aggregate(result[,c("EBLUP", "GREG", "Synth",   "non_domain")],  by=lis
56
```

50:1 (Top Level) R Script

Console ~/

```
Linear mixed-effects model fit by REML
 Data: amigo1
       AIC      BIC    logLik
  8104.348 8140.741 -4046.174

Random effects:
 Formula: ~1 | judet
        (Intercept)  Residual
StdDev:    5.270207 0.5454103

Fixed effects: absent ~ edu_inf + mediu_r + gr_5
                Value  Std.Error   DF   t-value p-value
(Intercept)  2.2439958 0.27565918 2596  8.140472       0
edu_inf      0.1191729 0.00542043 2596 21.985856       0
mediu_r      0.0080927 0.00176590 2596  4.582749       0
gr_5        -0.0546386 0.00795100 2596 -6.871915       0
 Correlation:
        (Intr) edu_nf medi_r
edu_inf  0.003
mediu_r -0.198  0.584
gr_5    -0.347 -0.704 -0.545

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.360346e+01 -9.692781e-04 -6.778054e-04 -5.373032e-04  1.666872e+01

Number of Observations: 3187
Number of Groups: 588
>
```

Workspace | History

Load | Save | Import Dataset | Clear All

Data

| | |
|---|---|
| amigo | 76986 obs. of 15 variables |
| amigo1 | 3187 obs. of 15 variables |
| amigo1_M | 3187 obs. of 14 variables |
| amigo1_absent | 3549 obs. of 15 variables |
| amigo2 | 42 obs. of 15 variables |
| amigo_judete | 42 obs. of 22 variables |
| amigo_judete_mean | 42 obs. of 23 variables |
| amigo_judete_sum | 42 obs. of 22 variables |
| d.data | 42 obs. of 13 variables |
| estimatori | 42 obs. of 5 variables |
| medii | 42 obs. of 4 variables |
| result | 42 obs. of 38 variables |

Files | Plots | Packages | Help

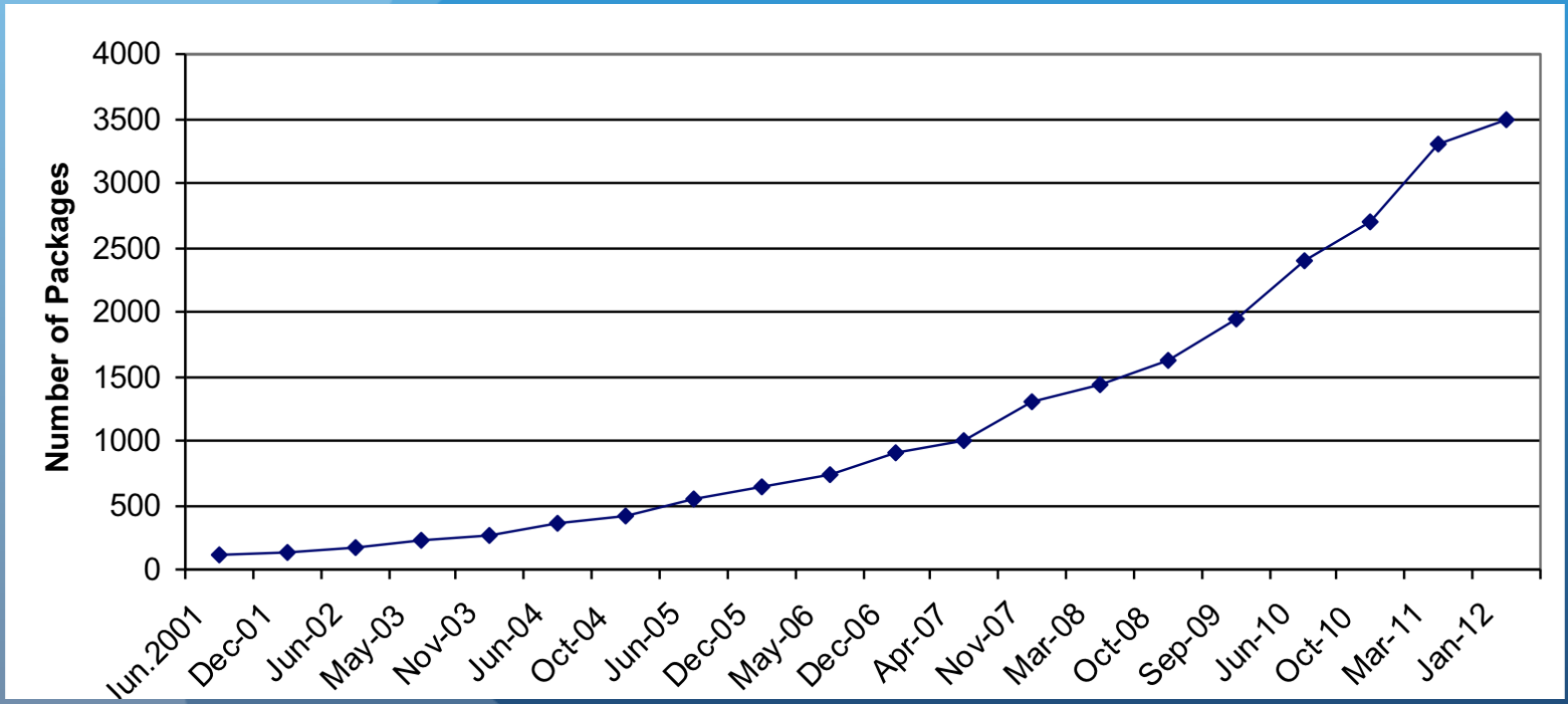
Install Packages | Check for Updates

| | | |
|---|---|---|
| ☐ | abind | Combine multi-dimensional arrays |
| ☐ | aplpack | Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, and some slider functions |
| ☐ | boot | Bootstrap Functions (originally by Angelo Canty for S) |
| ☑ | car | Companion to Applied Regression |
| ☐ | class | Functions for Classification |
| ☐ | cluster | Cluster Analysis Extended Rousseeuw et al. |
| ☐ | codetools | Code Analysis Tools for R |
| ☐ | colorspace | Color Space Manipulation |
| ☐ | compiler | The R Compiler Package |
| ☑ | datasets | The R Datasets Package |
| ☐ | Deducer | Deducer |
| ☐ | DeducerExtras | Additional dialogs and functions for Deducer |
| ☐ | DeducerMMR | A Deducer plugin for moderated multiple regressions and simple slopes analysis |
| ☐ | DeducerPlugInScaling | Reliability and factor analysis plugin |
| ☐ | DeducerSpatial | Deducer for spatial data analysis |
| ☐ | dichromat | Color schemes for dichromats |
| ☐ | digest | Create cryptographic hash digests of R objects |
| ☐ | e1071 | Misc Functions of the Department of Statistics (e1071), TU Wien |
| ☐ | effects | Effect Displays for Linear, Generalized Linear, Multinomial-Logit, Proportional-Odds Logit Models and mixed-effects models |
| ☐ | foreign | Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ... |
| ☐ | ggplot2 | An implementation of the Grammar of Graphics |

Start | Inbox - Microsof... | GEBA 2012 - An... | Downloads | Microsoft Power... | RStudio | R Commander | 12:13 PM

# R Packages

- Extended through packages, user-created add-on programs
- Every package is a research project that is reviewed at academic level




Sources: *http://r4stats.com/articles/popularity/ and http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Fox.pdf*

# SWOT

## Strengths

- ❑Freeware, open-source
- ❑Easy to install and configure on various operating systems
- ❑Linked with the way statisticians think and work
- ❑Various procedures not available in SPSS and SAS
- ❑The freedom to teach with real-world examples from outside organizations
- ❑The flexibility to mix-and-match models, scripts and packages for the best results

## Weaknesses

- ❑Data collection should be available from other tools; MySQL or PostgreSQL are popular among R users for this purpose
- ❑Guided Analytics not available
- ❑The help files and the vignettes for packages are written for relatively advanced users
- ❑R is not very user friendly and it needs basic knowledge of programming language

## Opportunities

- ❑Share new techniques with other R users around the world via online community
- ❑Re-use and reproduce new discovered techniques on analytic operations that the user is going to perform – this is difficult in SAS or SPSS
- ❑Very large area of use - statistics, journalism, mapping, finance, forecasting, social networking, drug development, computational biology, life sciences and many more

## Threats

- ❑Harder to learn than other similar software due to the fact that it has more types of data structures than the data set
- ❑It is necessary for the user to carry out the macro language of R and to control the management of the output; SPSS and SAS allow user to skip those issues until he needs them

# STATISTICAL ANALYSIS WITH R

## An example: The simple linear regression model

- $y_i = b_0 + b_1 x_i + e_i$

y = response variable

x = covariate

• Linear models are fit using R's model formulas.

# Estimating the parameters in simple linear regression

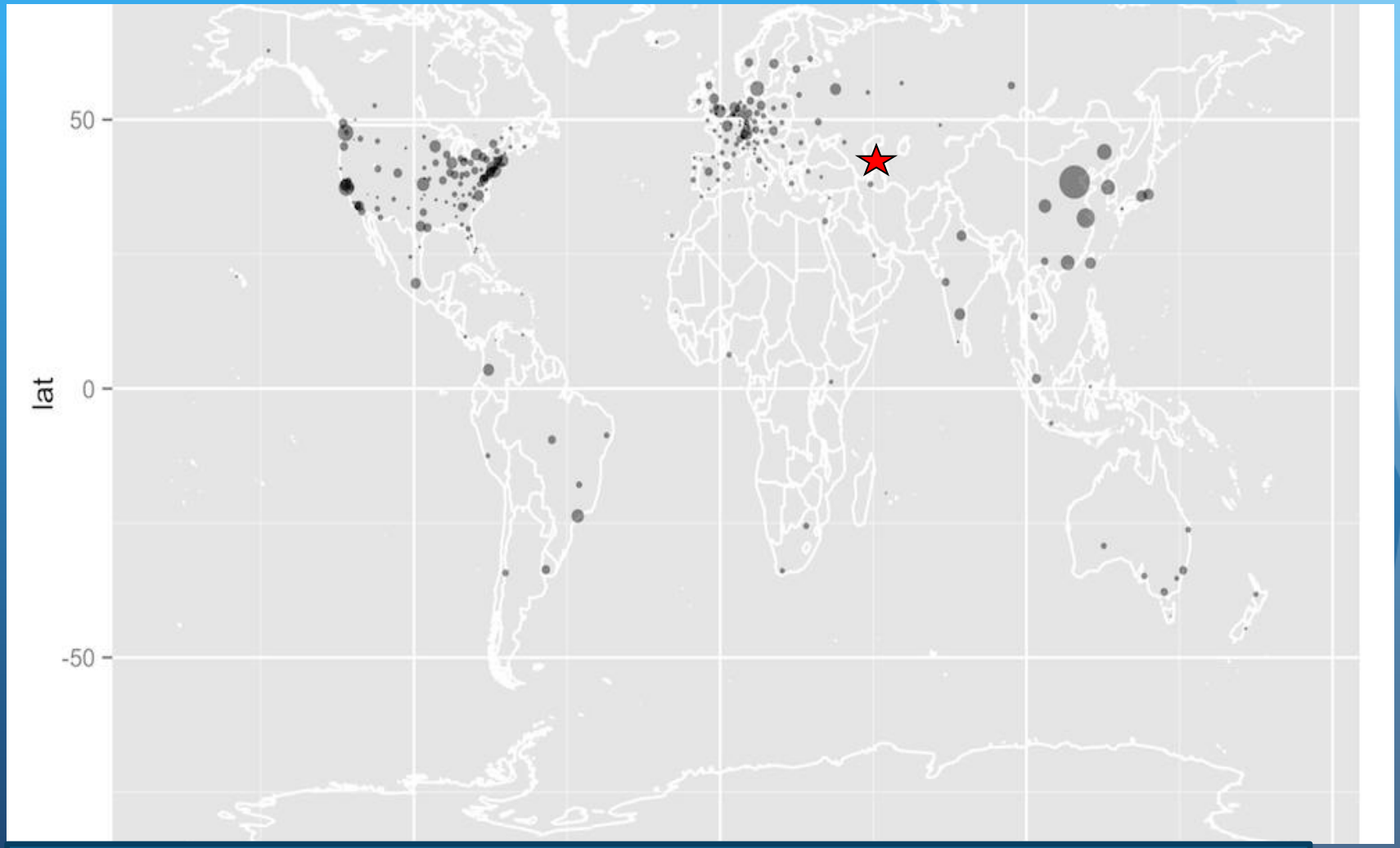
- The basic format for a formula is the ~ (tilde) is read “is modeled by” and is used to separate the response from the predictor(s).
- One goal when modeling is to “fit” the model by estimating the parameters based on the sample. For the regression model is used the method of least squares.
- ***To find the estimates, it is used the lm() function.*** The basic usage of lm function is of the form:

  lm(formula, data=…, subset=…)

# R output sample – lm function

```
> fit.lm <- lm (income ~ edu_level + occup)
> summary(fit.lm)

Call:
lm(formula = income ~ edu_level + occup)

Residuals:
    Min      1Q  Median      3Q     Max
-9370.2  -203.6  -107.4    55.2  9729.2

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 4.084e+02  1.800e+01   22.686  < 2e-16 ***
edu_level   1.715e+00  6.351e-04 2700.399  < 2e-16 ***
occup       2.608e-02  5.164e-03    5.051 4.64e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 562.1 on 3178 degrees of freedom
Multiple R-squared: 0.9996,    Adjusted R-squared: 0.9996
F-statistic: 3.669e+06 on 2 and 3178 DF,  p-value: < 2.2e-16
```

# R – global bRainstoRming




*Source: http://www.r-bloggers.com/where-in-the-world-is-r-and-rstudio/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+RBloggers+%28R+bloggers%29*

# Conclusion

- “R is the most powerful and flexible statistical programming language in the world”. - Norman Nie, co-founder of SPSS in the late 1960’s, currently, CEO and president of Revolution Analytics, a company that provides commercialized versions of R programs
- Official statistical systems using R: Italy, Austria, Australia, Canada
- Companies using R: Pfizer, Shell, Facebook, Google, Mozilla, Times, The New York Times, The Economist, NewScientist, Lloyd's, Bing, Johnson&Johnson

# Romanian new useRs

- Unfortunately, Romania is not on the current available users list worldwide, but it is not late to make this happen.
- Our country has very good and competitive computer scientists, which could become useRs.
- For the moment, there is a small group in the official statistic involved in small area estimation based on R technique.

# Thank you!

**If you would like to join Romanian team, please contact us at:**
**romanian.r.team@gmail.com**