

IX Jornadas de R



Informes Rmarkdown interactivos como resultado del análisis avanzado de datos ómicos



Tau Analytics

Data Science, Estadística y Bioinformática
info@tauanalytics.es
+34 687 79 38 61
C/ Catedrático Agustín Escardino, 9
Parc Científic de la Universitat de València

Reproducibilidad

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
<http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573>

Growth in a Time of Debt

By CARMEN M. REINHART AND KENNETH S. ROGOFF[✉]

In this paper, we exploit a new multi-country historical dataset on public (government) debt to search for a systemic relationship between high public debt levels, growth and inflation.¹ Our main result is that whereas the link between growth and debt seems relatively weak at “normal” debt levels, median growth rates for countries with public debt over roughly 90 percent of GDP are about one percent lower than otherwise; average (mean) growth rates are several percent lower. Surprisingly, the relationship between public debt and growth is remarkably

especially against the backdrop of graying populations and rising social insurance costs? Are sharply elevated public debts ultimately a manageable policy challenge?

Our approach here is decidedly empirical, taking advantage of a broad new historical dataset on public debt (in particular, central government debt) first presented in Carmen M. Reinhart and Kenneth S. Rogoff (2008, 2009b). Prior to this dataset, it was exceedingly difficult to get more than two or three decades of public debt data even for many rich countries, and

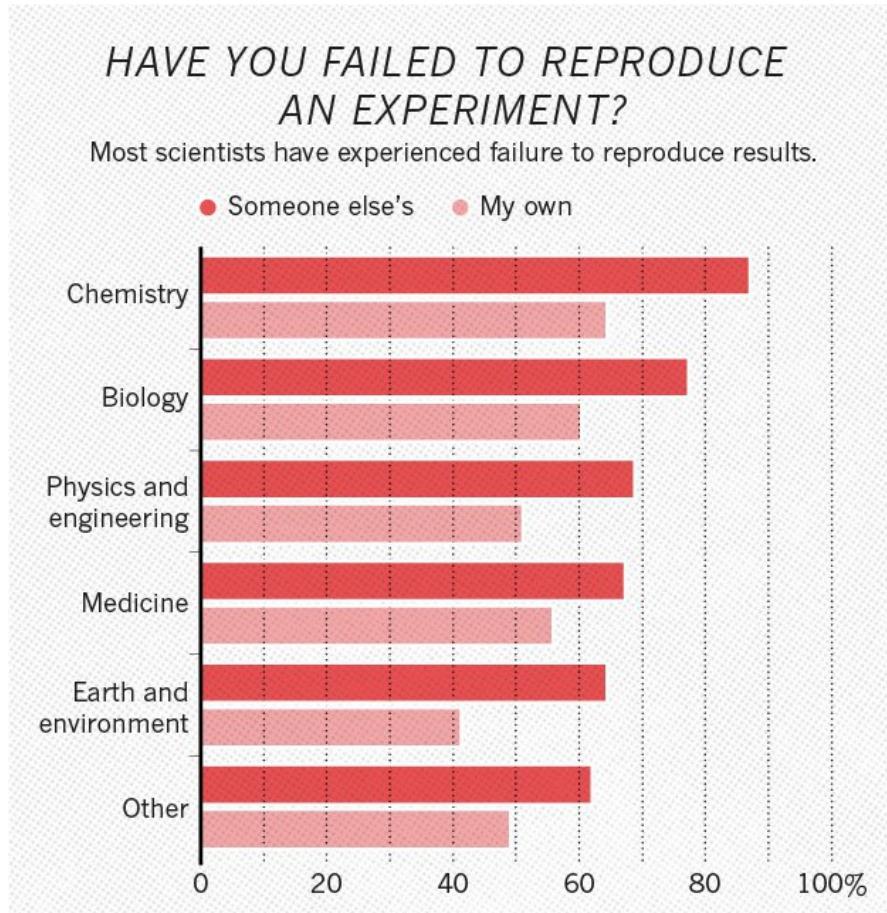
Growth in a Time of Debt

[CM Reinhart, KS Rogoff - American economic review, 2010 - pubs.aeaweb.org](#)

By Carmen M. Reinhart and Kenneth S. Rogoff* especially against the backdrop of graying populations and rising social insurance costs? Are sharply elevated public debts ultimately a manageable policy challenge? Our approach here is decidedly empirical, taking advantage ...

☆ 99 Citado por 3212 Artículos relacionados

Investigación y reproducibilidad



Nature 533, 452–454 (26 May 2016)

Crisis metodológica

Técnicas estadísticas utilizadas en proyectos y artículos

ANOVA

Test t

Chi-cuadrado

>95% de los artículos

Mann-Whitney

Correlación de Pearson



Ordenador analógico (1945)



Balanza de precisión (1889)

Datos ómicos



El análisis de datos ómicos se caracteriza por la presencia de un muy elevado número de variables (**p>>n**)

Genómica (500000 SNPs), transcriptómica (5000 miRNAs), epigenómica (900000 CpGs)...

Inconvenientes de los análisis clásicos

1. Comparaciones múltiples →  tasa FP → Corrección p-values →  tasa FN
2. No hay control de las variables confusión o interacciones entre variables
3. Difícil interpretación e integración de los resultados

Alternativas de análisis para datos ómicos

Para analizar este tipo de datos existen dos alternativas:

1. Darle la vuelta al problema y utilizar cada gen (miRNA, proteína, etc.) como variable respuesta

$$\text{expresión}_{gen_i} = \beta_0 + \beta_1 * grupoB + \beta_2 * batch2 + \varepsilon$$

Diferencia entre los grupos A y B corrigiendo por el batch effect

- La idea es similar a la de hacer test t, pero ahora podemos añadir otras variables de control al análisis

2. Utilizar métodos modernos que nos permitan analizar

$$P(\text{enfermedad}) \sim \beta_0 + \beta_i * gen_i + \varepsilon$$

ALERT

$p >> n$

¿df? → Métodos de penalización

Otros modelos de regresión

- ➡ Binomial negativa (recuentos con sobredispersión)
RNAseq
- ➡ Beta (variables continuas delimitadas en un rango)
Arrays de metilación
- ➡ Logística multinomial (variables categóricas)
- ➡ Ordinal (variables ordinales)
- ➡ Gamma (variables continuas estrictamente positivas)
- ➡ Rank regression (valores extremos en la respuesta)

...

Caso práctico

- Patología: Alzheimer
- Dos variantes: Genética (EAG) o Esporádica (EAE)
- 900.000 datos (CpGs) por individuo
- 5 individuos EAG y 5 individuos en EAE ($p>>n$)

¿Existen diferencias en el perfil epigenético?

Análisis estadísticos



Resultados
Informe interactivo



Rmarkdown

Una apuesta por la transparencia y la reproducibilidad

¿Qué es Rmarkdown?

Lenguaje de marcado



* _ __ # ##



Convierte el lenguaje en
HTML válido



Aplicación de
formato

Simplificación de la programación

HTML

```
<b>data</b>  
<strong>data</strong>
```

markdown

 data
data

Output

data

```
<i>data</i>  
<em>data</em>
```

 data
data

data

```
<h1>Heading</h1>
```

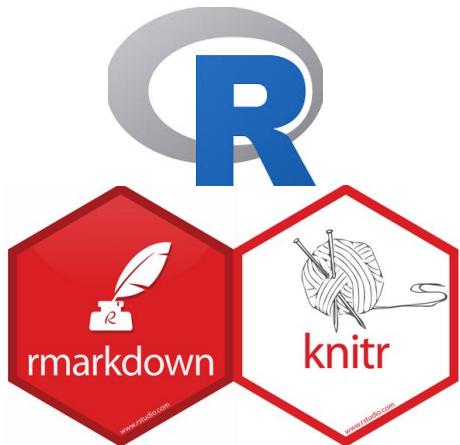
#Heading

Heading

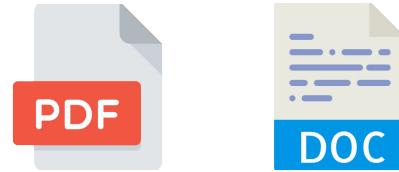
9 caracteres → 1



Funcionamiento



Outputs



Code chunks

```
278
279  ## Heatmap
280
281 En el heatmap obtenido a partir de una muestra aleatoria de 5000 CpGs, no se revela un claro patrón de diferenciación. Aunque el grupo Sporadic FTD's disease - Tau protein_ parece tener niveles mayoritariamente hipermetilados, las muestras BK1751 y BK1681 tiene un patrón de metilación parecido a los controles (hipometilados).
282
283 ````{r heatmap4, echo=FALSE, message=FALSE, warning=FALSE,fig.align='center',fig.width=10,fig.height=5}
284 NMF::ahexmap(t(scale(datos_f$datos_ford$Grupo %in% c('G1','G4_DFT_esp')),c(sample(26:841036, 5000)))),
285     annCol = data.frame(Grupo=factor(datos_ford$defin[datos_ford$Grupo %in% c('G1','G4_DFT_esp')]),
286     `PostMortem delay (h)`=datos_ford$postmortem_delay_h[datos_ford$Grupo %in% c('G1','G4_DFT_esp')]),
287     main='',color=c("green", "black","red"), scale="row", breaks=0, Colv = NA,
288     labCol = datos_ford$codigo[datos_ford$Grupo %in% c('G1','G4_DFT_esp')])
289 ...
290
291  ## t-SNE
```

Ventajas

extensión (2014)

Outputs

1. Nivel de programación requerido básico

PDF

DOC

2. Bajo coste de tiempo de elaboración

rmarkdown

knitr

3. Versátil e interactivo

Code chunks

4. Reproducibilidad de los resultados

```
278
279 <### Heatmap
280
281 En el heatmap obtenido a partir de una muestra aleatoria de 5000 CpGs, no se revela un claro patrón de diferenciación. Aunque el grupo „Sporadic FTD's disease...“ - Tau
protein_ parece tener niveles mayoritariamente hipermetilados, las muestras BK1751 y BK1681 tiene un patrón de metilación parecido a los controles (hipometilados).
282
283 + """{r heatmap4, echo=FALSE, message=FALSE, warning=FALSE,fig.align='center',fig.width=10,fig.height=5}
284 NMF::ahemap(t$scale(datos_ford$Grupo %in% c('G1','G4_DFT_esp')),c(sample(26:841036, 5000)))
285     annCol = data.frame(Grupo=factor(datos_ford$ford$Grupo %in% c('G1','G4_DFT_esp'))),
286     "PostMortem delay (h)"=datos_ford$postmortem_delay_h[datos_ford$Grupo %in% c('G1','G4_DFT_esp')]),
287     main='',color=c("green", "black","red"), scale="row", breaks=0, Colv = NA,
288     labCol = datos_ford$codigo[datos_ford$Grupo %in% c('G1','G4_DFT_esp')])
```

CONTACTO



Tau Analytics

Data Science, Estadística y Bioinformática



info@tauanalytics.es

+34 687 79 38 61

Parc Científic de la
Universitat de València
C/Catedrático Agustín Escardino, 9