

The model comparison approach in teaching statistics

The R2STATS and AtelierR GUIs for R

Yvonnick Noël

Department of Psychology, University of Brittany, Rennes

"Deuxièmes Rencontres R", Lyon June 2013

Sommaire

- 1 Typical statistical problems in the social sciences
- 2 A simple example: Comparing means
 - The classical, test-oriented approach
 - The model comparison approach
 - The R2STATS GUI for `glm()`
- 3 Future developments

The common statistical toolbox for psychologists

- Colleagues generally want their undergraduate students to be able to:
 - 1 compare proportions (z-test, χ^2),
 - 2 compare categorical distributions or test for independence between categorical variables (χ^2)
 - 3 compare means (one and two sample Student t , Fisher F),
 - 4 compare variances (Fisher F , Levene),
 - 5 model linear dependencies (correlation, regression).
- That's most of it...

The problem oriented approach

- Colleagues' demands to the stats teacher are often **problem-oriented** (e.g. compare means).
- Standard statistical packages usually offers this **hardwired in their menus**. This is also straightforward in R (`t.test()`, `chisq.test()`, etc).
- An **apparent benefit** of this strategy is that students may be quickly "autonomous".
- But...

Age of onset of schizophrenia by gender

```
> data(schizophrenia)
> head(schizophrenia)
  age gender
1  20 female
2  30 female
3  21 female
4  23 female
5  30 female
6  25 female
...
```

Age of onset of schizophrenia by gender

```
> t.test(age~gender,data=schizophrenia)
```

Welch Two Sample t-test

data: age by gender

t = 4.7989, df = 166.077, p-value = 3.533e-06

alt. hypo.: true difference in means is not equal to 0

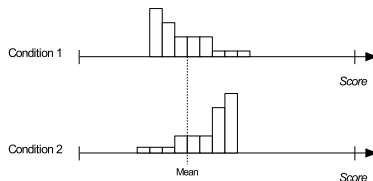
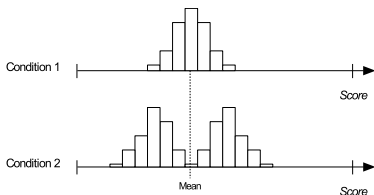
95 percent confidence interval: 3.861288 9.259260

sample estimates:

mean in group female	mean in group male
30.47475	23.91447

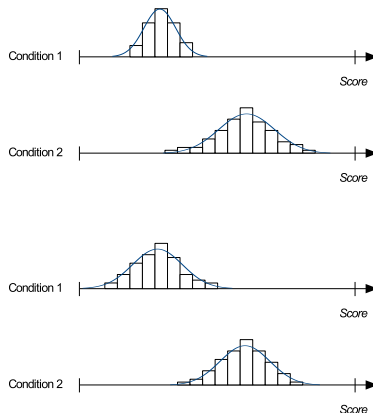
Comparing distributions

- Comparing experimental conditions, supposed to induce different behaviors, statistically means **comparing distributions**.
- In many cases, this is simplified into a **mean-comparison problem**.
- But... What if comparing means is **meaningless** in the first place?

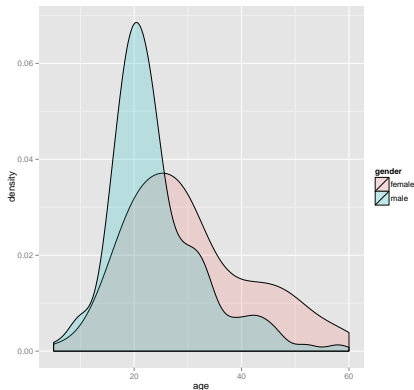


Comparing means

- **Comparing means** has the advantage of simplicity, in terms of interpretation. But **not always meaningful** (even in the unimodal Gaussian case).
- This makes sense under a **symmetric, unimodal** and **homoscedastic** Gaussian model, as only the expectation changes between groups.



Age of onset of schizophrenia: Density estimates



Data are times (left-bounded, continuous). The **population distributions** are most probably:

- asymmetric,
- heteroscedastic.

Questions

- Is it meaningful to compare means when **asymmetric heteroscedastic** distributions are suspected?
- **Proposal**: Yes, provided:
 - a valid **distribution model** with these attributes is available,
 - with a **mean parameter** (for which the empirical mean is the maximum likelihood estimator),
 - **only this parameter** is assumed to change from one group to the other.

What we really need

- What we need is a **model-oriented** way of thinking: What is the distribution underlying my data?
- Ideally, the answer should result from an hypothesis on the **data generation mechanism**.
- Once a distribution model is chosen, descriptive statistics (means, variances, proportions, etc.) **follow as consequences** (i.e. parameters in the model).

Times to onset

- Ages of onset of schizophrenia by gender are **times-to-event** variables (T_j , $j = 1, 2$).
- A candidate model is a conditional **Gamma distribution**
 $T_j \sim \Gamma(s, \mu_j)$ (note: Mean parameterization):

$$f_j(t|s, \mu_j) = \frac{1}{\Gamma(s)} \left(\frac{s}{\mu_j} \right)^s t^{s-1} e^{-\frac{st}{\mu_j}}$$

- This is very much like assuming that we are waiting for **the same fixed number** (s) of hidden steps before onset, occurring at **different rates** ($\lambda_j = s/\mu_j$) in both groups.

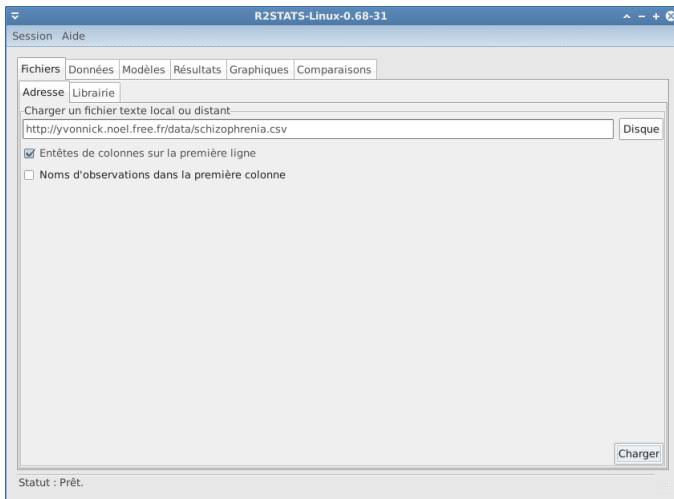
The model comparison workflow in R

- The problem now amounts to:
 - **test goodness-of-fit** of a Gamma model,
 - **compare** the two-means Gamma model with its constrained version ($\mu_j = \mu, \forall j$).

- This model comparison approach is **natural in R**:

```
data(schizophrenia)
M0 = glm(age~1,family=Gamma,data=schizophrenia)
M1 = glm(age~1+gender,family=Gamma,data=schizophrenia)
anova(M0,M1, test="F")
```

Loading the data



Examining the data

The screenshot shows the R2STATS-Linux-0.68-31 application window. The 'Données' (Data) tab is active, displaying a table titled 'schizophrenia'. The table has three columns: 'age' and 'gender'. The data is as follows:

	age	gender
1	20	female
2	30	female
3	21	female
4	23	female
5	30	female
6	25	female
7	13	female
8	19	female
9	16	female
10	25	female
11	20	female
12	25	female
13	27	female
14	43	female

At the bottom of the window, the status bar indicates 'Statut : Prêt.' (Status: Ready).

Model definition: The Gaussian homoscedastic model

R2STATS-Linux-0.68-31

Session Aide

Fichiers Données **Modèles** Résultats Graphiques Comparaisons

Tableau des données

schizophrenia

Variables Type

age	N
gender	F

Résumé de variable

Attribut	Valeur
female	99
male	152
Manquantes	0
Total	251

Définition de modèle

Nom du modèle M1

Variables dépendantes

age

Ajouter Effacer

Variables indépendantes

gender

Ajouter + : * - () Fixée +1 +0 | (1.) (.) / Effacer

Loi de distribution Normale

Fonction de lien Identique

Variable de pondération Aucune

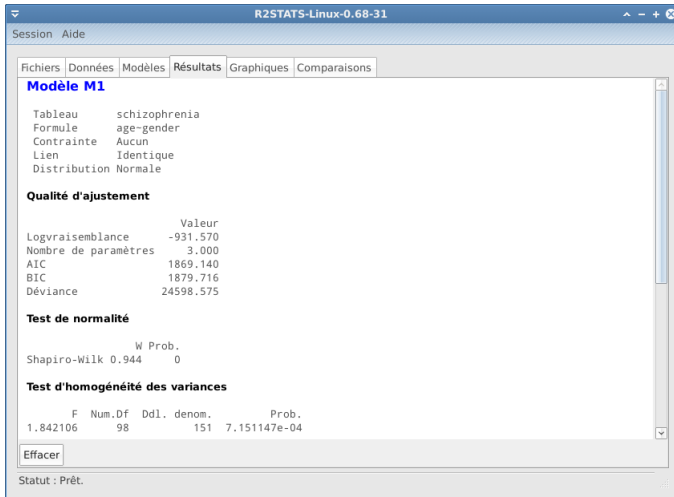
Facteur de contrainte Aucun

Sélection d'observation

Estimation

Statut : Prêt.

Goodness-of-fit



The screenshot shows the R2STATS-Linux-0.68-31 application window. The 'Résultats' tab is selected, displaying the following information for 'Modèle M1':

Modèle M1

Tableau	schizophrenia
Formule	age-gender
Contrainte	Aucun
Lien	Identique
Distribution	Normale

Qualité d'ajustement

	Valeur
Logvraisemblance	-931.570
Nombre de paramètres	3.000
AIC	1869.140
BIC	1879.716
Déviance	24598.575

Test de normalité

	W Prob.
Shapiro-Wilk	0.944 0

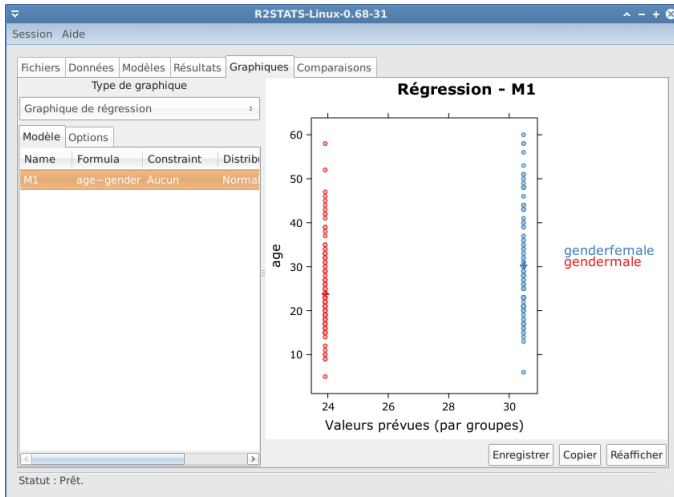
Test d'homogénéité des variances

F	Num.Df	Ddl. denom.	Prob.
1.842106	98	151	7.151147e-04

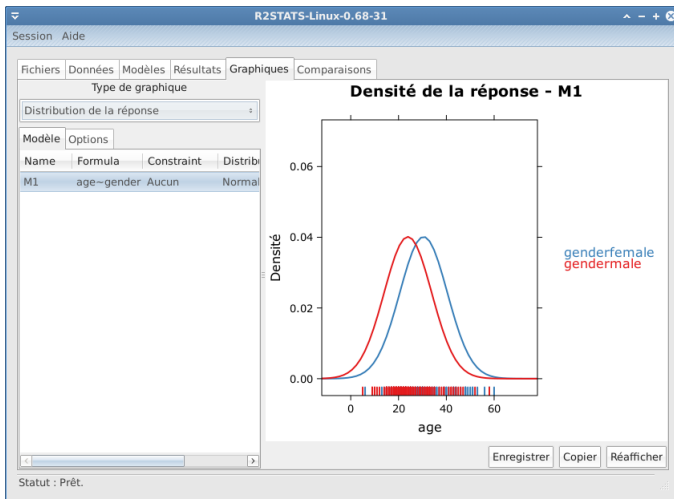
Effacer

Statut : Prêt.

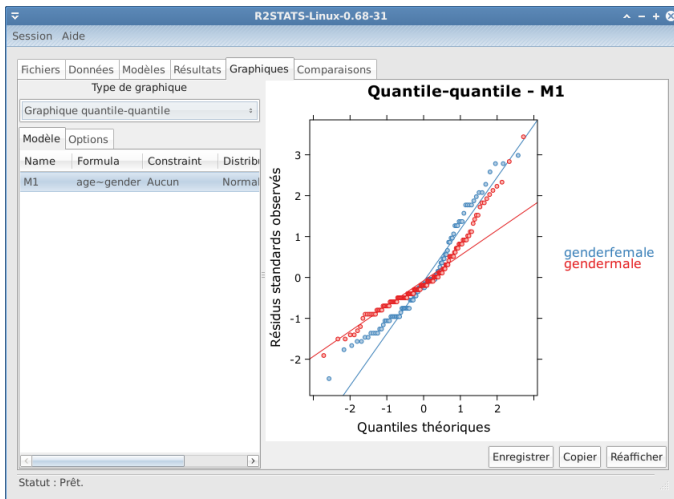
Plots: Regression



Plots: Response distribution



Plots: Quantile-quantile plot



Model definition: The fixed shape Gamma model

R2STATS-Linux-0.68-31

Session Aide

Fichiers Données Modèles Résultats Graphiques Comparaisons

Tableau des données

schizophrenia

Variables Type

age	N
gender	F

Résumé de variable

Attribut	Valeur
female	99
male	152
Manquantes	0
Total	251

Définition de modèle

Nom du modèle M1bis

Variables dépendantes

age

Ajouter Effacer

Variables indépendantes

gender

Ajouter + : * - () Fixée +1 +0 | (1.) (.) / Effacer

Loi de distribution

Gamma

Fonction de lien

Inverse

Variable de pondération

Aucune

Facteur de contrainte

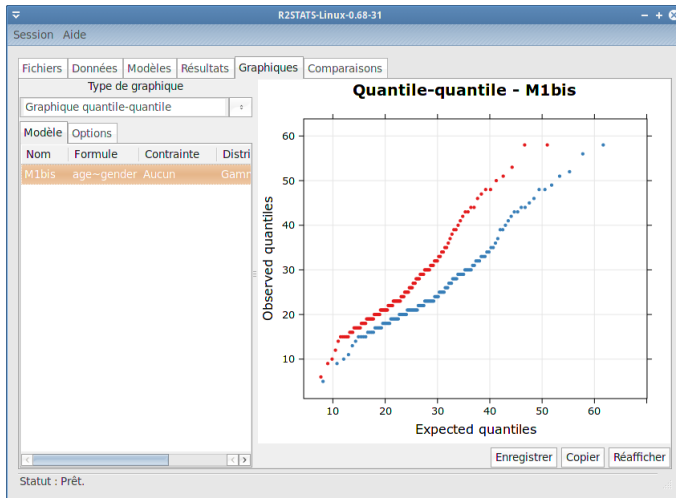
Aucun

Sélection d'observation

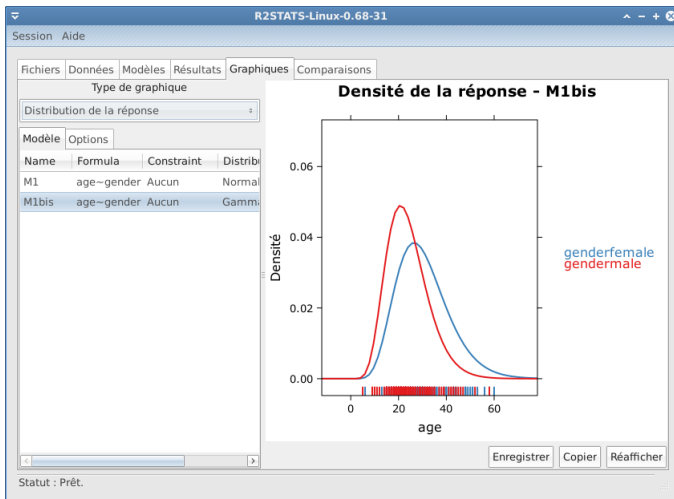
Estimation

Statut : Prêt.

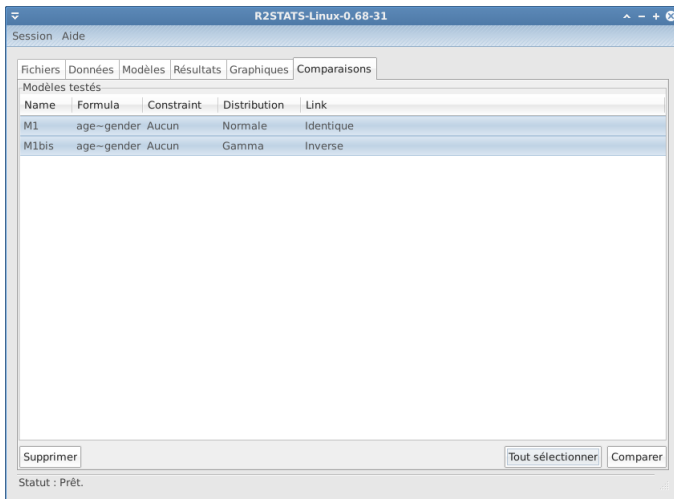
Plot: Quantile-quantile plots



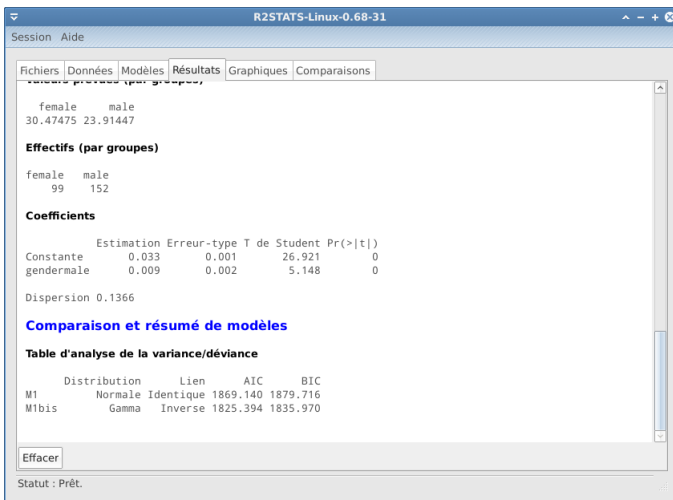
Plots: Response distribution



Model comparison



Model comparison



The screenshot shows the R2STATS-Linux-0.68-31 GUI. The 'Résultats' tab is selected, displaying the following output:

```
female    male
30.47475 23.91447
```

Effectifs (par groupes)

```
female    male
   99     152
```

Coefficients

	Estimation	Erreur-type	T de Student	Pr(> t)
Constante	0.033	0.001	26.921	0
gendermale	0.009	0.002	5.148	0

Dispersion 0.1366

Comparaison et résumé de modèles

Table d'analyse de la variance/déviance

	Distribution	Lien	AIC	BIC
M1	Normale	Identique	1869.140	1879.716
M1bis	Gamma	Inverse	1825.394	1835.970

Effacer

Statut : Prêt.

Constrained Gamma model

R2STATS-Linux-0.68-31

Session Aide

Fichiers Données **Modèles** Résultats Graphiques Comparaisons

Tableau des données

schizophrenia

Variables	Type
age	N
gender	F

Résumé de variable

Attribut	Valeur
female	99
male	152
Manquantes	0
Total	251

Définition de modèle

Nom du modèle M0

Variables dépendantes

age

Ajouter Effacer

Variables indépendantes

gender

Ajouter + : * - () Fixée +1 +0 | (1.) (.1) / Effacer

Loi de distribution Fonction de lien Variable de pondération Facteur de contrainte

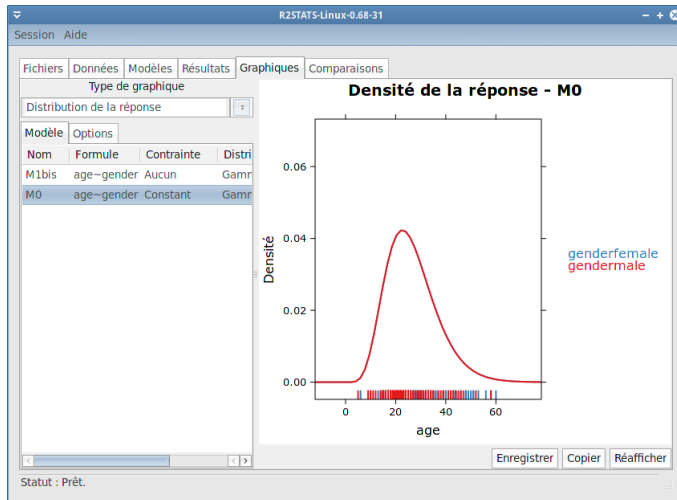
Gamma Inverse Aucune Constant

Sélection d'observation

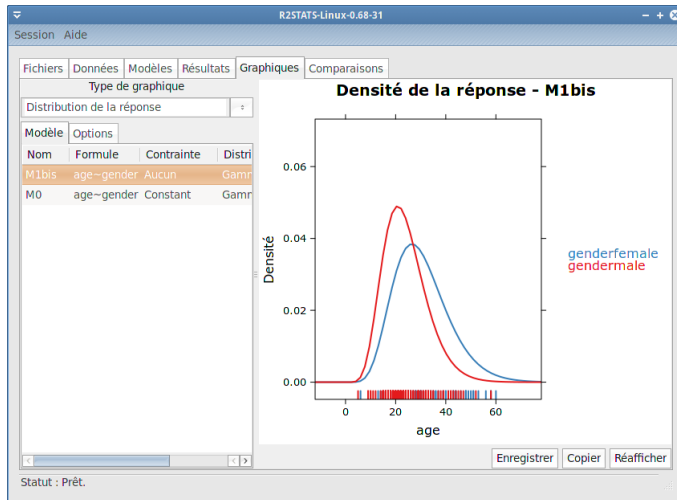
Estimation

Statut : Prêt.

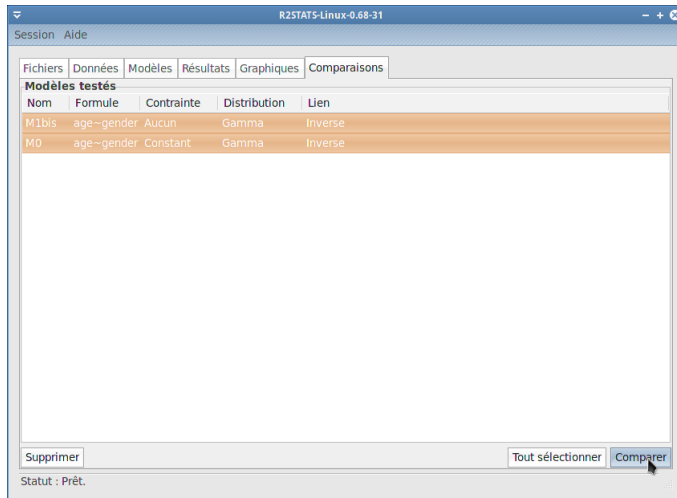
Plot: Response distribution under the constant Gamma model



Plots: Response distribution under the fixed shape Gamma model



Model comparison



Model comparison

R2STATS-Linux-0.68-31

Session Aide

Fichiers Données Modèles Résultats Graphiques Comparaisons

```

female    male
26.50199  26.50199

Effectifs (par groupes)

female    male
  99      152

Coefficients

              Estimation Erreur-type T de Student Pr(>|t|)
Constante      0.038      0.001      40.269      0

Dispersion 0.1548

Comparaison et résumé de modèles

Table d'analyse de la variance/déviance

      Formule Contrainte Distribution Lien
M0      age-gender Constant      Gamma Inverse
M1bis   age-gender   Aucun      Gamma Inverse

      Ddl rés. Dév. res. Ddl diff. Rap.Vr.      F Pr(>F)      AIC BIC Expl.(%)
M0      250    36.912
M1bis   249    33.337      1    3.5754    26.174 6.2341e-07 1825.4 1836.0 0.096863
    
```

Effacer

Statut : Prêt.

Conclusions

- The **problem-oriented** approach in stats teaching leads to **bad practices**, either bad applications (e.g. t.test) or bad restrictions (e.g. no comparison of means).
- By contrast, the model-oriented approach:
 - makes it possible to deal with unusual situations (compare means in the assymetric heteroscedastic case),
 - reduces the gap between substantial theory and statistics: Not so much "Statistics applied to psychology" but "**Statistical psychology**" (hypothesis on a data generation mechanism).

R2STATS on the Web (I)

The screenshot shows the R2-GLMM web application running in a Mozilla Firefox browser. The browser's address bar shows the URL 'R2-GLMM'. The application's navigation bar includes links for 'Home', 'R2-LRN', 'R2-PROBCALC', 'R2-BAYES', 'R2-GLM(M)', and 'R2-SEM'. The main heading is 'R2-GLM A Web GUI for fitting GLM and GLMM'. Below this, there are tabs for 'Files', 'Data', 'Models', 'Results', 'Plots', and 'Comparisons'. The 'Data' tab is active, displaying a table of 10 entries. The table has columns for 'Treat', 'Prewt', and 'Postwt'. The 'Treat' column contains the value 'Cont' for all entries. The 'Prewt' and 'Postwt' columns contain numerical values. At the bottom of the table, it says 'Showing 1 to 10 of 72 entries'. There are 'Previous' and 'Next' navigation links. The footer of the application states '© Yvonnick Noel, University of Brittany, Rennes 2, 2013'.

	Treat	Prewt	Postwt
1	Cont	81	80
2	Cont	89	80
3	Cont	92	86
4	Cont	74	86
5	Cont	78	76
6	Cont	88	78
7	Cont	87	75
8	Cont	75	87
9	Cont	81	74
10	Cont	78	85

R2STATS on the Web (II)

R2-GLMM - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils Aide

R2-GLMM

Home R2-LRN R2-PROBCALC R2-BAYES R2-GLM(M) R2-SEM

Files Data Models Results Plots Comparisons

Data

Anorexia

Variables

Prewt
Postwt
Treat

Summary

Attribute	Value
Control	29
CBT	26
FT	15

Model

Name M1 Last models

Dependent variable

Postwt

Add Clear

Independent variables

Prewt*Treat

Add + * ^ () Offset +1 +0 | (1.) (. / Clear

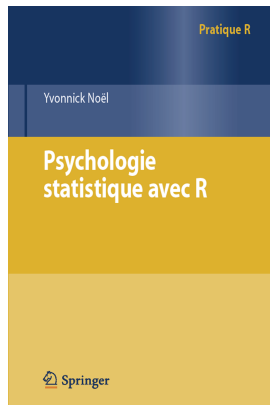
Distribution Link Weights Constraint

Gaussian Identity None None

Selection

Type a valid R expression...

Bibliography



- Binomial, multinomial and Gaussian models
- A model comparison approach, both in a Fisherian and Bayesian perspective
- A number of real examples from psychological research
- The R2STATS and AtelierR packages.