II. Data exploration & Statistical analyses

Florent Manzi
Feb 2020

Disease Eco-Evo Lab

In this course we will see:

- How to **explore your data**, verify your model assumptions
- Build simple (lm) and generalized (glm) **linear models**
- Perform simple **ANOVAs** (analyses of variance) and **post-hoc tests**
- How to perform **model comparison** (stepwise regression, using AIC)

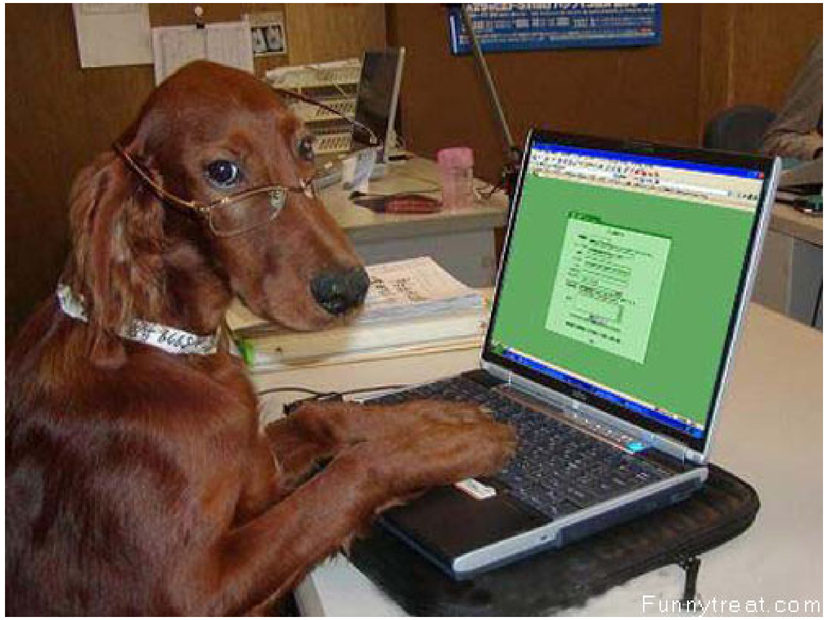
In this course we will see:

- How to **explore your data**, verify your model assumptions
- Build simple (lm) and generalized (glm) **linear models**
- Perform simple **ANOVAs** (analyses of variance) and **post-hoc tests**
- How to perform **model comparison** (stepwise regression, using AIC)

- Plot your data using **ggplot2** and the **Hmisc** package
- Useful tips to organize your layout, change the legends, etc.
- Export your plot in **high resolution** (for publication / presentation...)

Mind the dogs !

The **Listen** Dog



The **Do** Dog

Is your data well organized?

E82

	Exposure	Food	Clone	Replicate	Total_offspring	Survived	Age_death	Infected	Spore_count
2	Control	Scenedesmus	AMME_12	1	17	1	35	0	0
3	Control	Scenedesmus	AMME_12	2	11	1	35	0	0
4	Control	Scenedesmus	AMME_12	3	13	1	35	0	0
5	Control	Scenedesmus	AMME_12	4	12	1	35	0	0
6	Control	Scenedesmus	AMME_12	5	19	1	35	0	0
7	Control	Scenedesmus	AMME_12	6	16	1	35	0	0
8	Control	Scenedesmus	AMME_12	7	12	1	35	0	0
9	Control	Scenedesmus	AMME_12	8	2	0	7	0	0
10	Control	Scenedesmus	AMME_12	9	9	1	35	0	0
11	Control	Scenedesmus	AMME_12	10	11	1	35	0	0
12	Control	Scenedesmus	AMME_12	11	12	1	35	0	0
13	Control	Scenedesmus	AMME_12	12	9	1	35	0	0
14	Control	Scenedesmus	AMME_12	13	11	1	35	0	0
15	Control	Scenedesmus	AMME_12	14	11	1	35	0	0
16	Control	Scenedesmus	AMME_12	15	13	1	35	0	0
17	Control	Scenedesmus	AMME_12	16	14	1	35	0	0
18	Control	Scenedesmus	AMME_12	17	17	1	35	0	0
19	Control	Scenedesmus	AMME_12	18	18	1	35	0	0
20	Control	Scenedesmus	AMME_12	19	6	0	18	0	0
21	Control	Scenedesmus	AMME_12	20	13	1	35	0	0
22	Control	Scenedesmus	MUGG_23	1	12	1	35	0	0
23	Control	Scenedesmus	MUGG_23	2	4	0	6	0	0

What you control

What you measure

Is your data well organized?

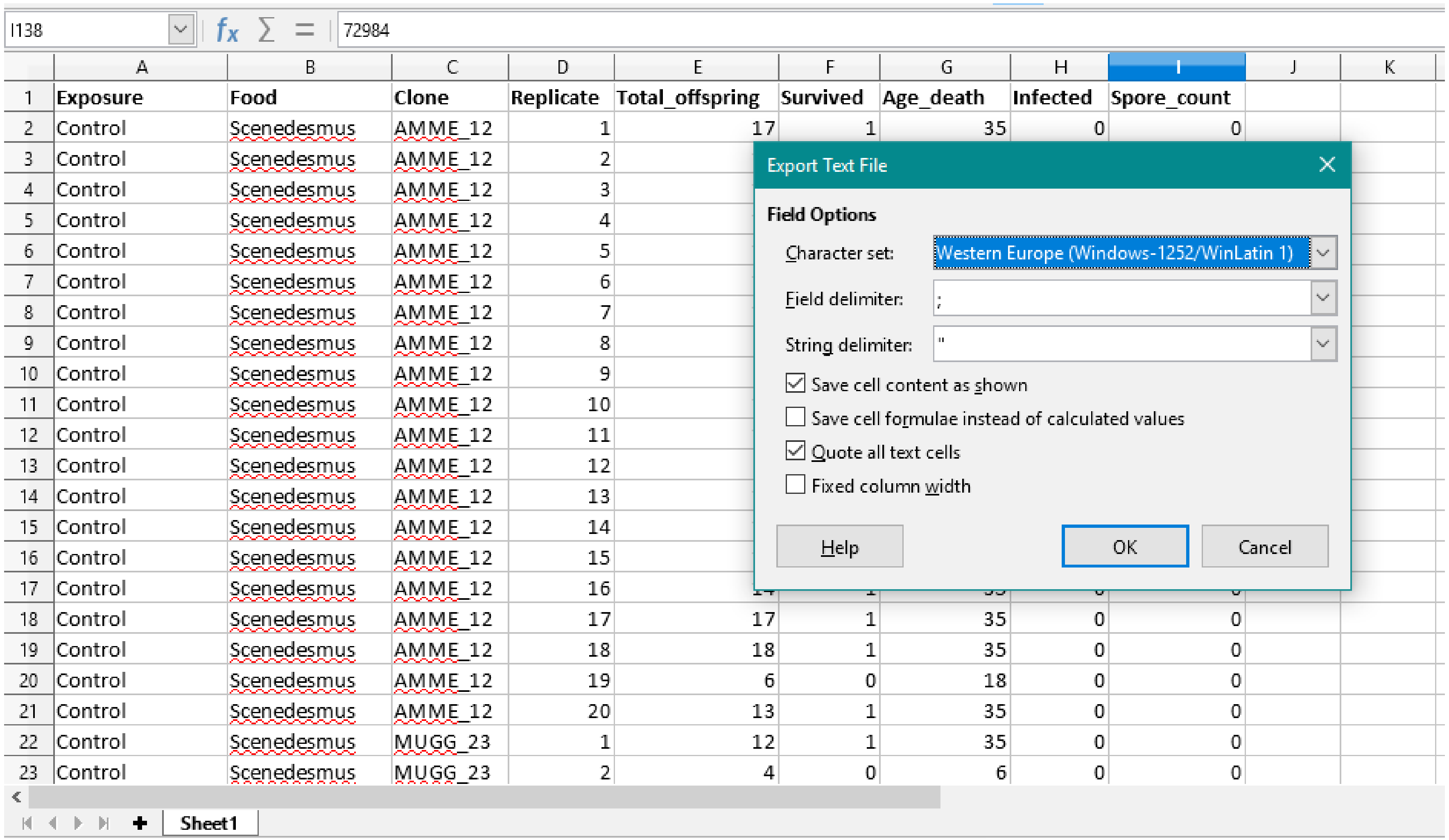
E82

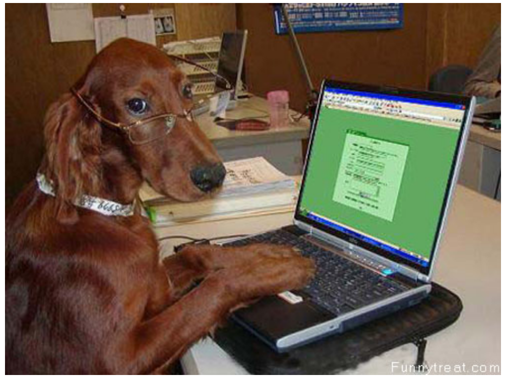
	A	B	C	D	E	F	G	H	I	J
1	Exposure	Food	Clone	Replicate	Total_offspring	Survived	Age_death	Infected	Spore_count	
2	Control	Scenedesmus	AMME_12	1	17	1	35	0	0	
3	Control	Scenedesmus	AMME_12	2	11	1	35	0	0	
4	Control	Scenedesmus	AMME_12	3	13	1	35	0	0	
5	Control	Scenedesmus	AMME_12	4	12	1	35	0	0	
6	Control	Scenedesmus	AMME_12	5	19	1	35	0	0	
7	Control	Scenedesmus	AMME_12	6	16	1	35	0	0	
8	Control	Scenedesmus	AMME_12	7	12	1	35	0	0	
9	Control	Scenedesmus	AMME_12	8	2	0	7	0	0	
10	Control	Scenedesmus	AMME_12	9	9	1	35	0	0	
11	Control	Scenedesmus	AMME_12	10	11	1	35	0	0	
12	Control	Scenedesmus	AMME_12	11	12	1	35	0	0	
13	Control	Scenedesmus	AMME_12	12	9	1	35	0	0	
14	Control	Scenedesmus	AMME_12	13	11	1	35	0	0	
15	Control	Scenedesmus	AMME_12	14	11	1	35	0	0	
16	Control	Scenedesmus	AMME_12	15	13	1	35	0	0	
17	Control	Scenedesmus	AMME_12	16	14	1	35	0	0	
18	Control	Scenedesmus	AMME_12	17	17	1	35	0	0	
19	Control	Scenedesmus	AMME_12	18	18	1	35	0	0	
20	Control	Scenedesmus	AMME_12	19	6	0	18	0	0	
21	Control	Scenedesmus	AMME_12	20	13	1	35	0	0	
22	Control	Scenedesmus	MUGG_23	1	12	1	35	0	0	
23	Control	Scenedesmus	MUGG_23	2	4	0	6	0	0	

Explanatory variables (Factors)

Response variables (Continuous, Binomial...)

Export your data safely (as .csv)





The purpose of statistical analyses

- You need an objective, non-biased tool to confirm (or disprove) the differences that you THINK you can see on your plots

- Otherwise, the goal is basically the same as plotting your data: you want to use **explanatory variables** (your x-axis, facet, colours…) to explain **differences in your response variable** (your y-axis)

- Therefore, if you know what kind of plot you want to make out of your data, then **you already know which statistical analyses you want to run !**

"Your hypothesis, is your plot, is your model"

Kate Laskowski, ca. 2019

Factors and levels

- In this course we will only focus on explanatory variables that take the form of **Factors** (as opposed to continuous variables)

Factor	Treatment
Factor levels	- No pesticide - Low dose - High Dose

3

Variable	Dose
Values	0 … 1.57 … 5.43 … 9.67 …

∞

Factors and levels

- In laboratory experiments, you often apply **treatments** to your organisms of interest
- For instance, you could expose your organism to **different diets**, or incubate them at different **temperatures**
- If you organized your data correctly, then remember that your **factors should be placed as header** (either *Food*, or *Temperature*), which can contain **several levels** (groups of rows with the same name)

Factors and levels

	A	B	C	D	E	F	G	H	I	J
1	Exposure	Food	Clone	Replicate	Total_offspring	Survived	Age_death	Infected	Spore_count	
2	Control	Scenedesmus	AMME_12	1	17	1	35	0	0	
3	Control	Scenedesmus	AMME_12	2	11	1	35	0	0	
4	Control	Scenedesmus	AMME_12	3	13	1	35	0	0	
5	Control	Scenedesmus	AMME_12	4	12	1	35	0	0	
6	Control	Scenedesmus	AMME_12	5	19	1	35	0	0	
7	Control	Scenedesmus	AMME_12	6	16	1	35	0	0	
8	Control	Scenedesmus	AMME_12	7	12	1	35	0	0	
9	Control	Scenedesmus	AMME_12	8	2	0	7	0	0	
10	Control	Scenedesmus	AMME_12	9	9	1	35	0	0	
11	Control	Scenedesmus	AMME_12	10	11	1	35	0	0	
12	Control	Scenedesmus	MUGG_23	1	12	1	35	0	0	
13	Control	Scenedesmus	MUGG_23	2	4	0	6	0	0	
14	Control	Scenedesmus	MUGG_23	3	11	1	35	0	0	
15	Control	Scenedesmus	MUGG_23	4	9	1	35	0	0	
16	Control	Scenedesmus	MUGG_23	5	12	1	35	0	0	
17	Control	Scenedesmus	MUGG_23	6	13	1	35	0	0	
18	Control	Scenedesmus	MUGG_23	7	6	0	12	0	0	
19	Control	Scenedesmus	MUGG_23	8	12	1	35	0	0	
20	Control	Scenedesmus	MUGG_23	9	9	1	35	0	0	
21	Control	Scenedesmus	MUGG_23	10	9	1	35	0	0	
22	Control	Microcystis	AMME_12	1	2	0	28	0	0	
23	Control	Microcystis	AMME_12	2	3	1	35	0	0	

Sheet1

Factor

Levels

ANOVA vs. t-test

- t-tests can be performed when you want to **compare the mean of only two groups:**

  (example: one factor, two levels → the ‘**pills**’ group and the ‘**placebo**’ group)

- ANOVAS should be performed when you want to **compare the means of at least three groups:**

  (example 1: one factor, three levels → the ‘**pills**’, ‘**placebo**’ and **‘control’** group) | **One-Way ANOVA**

  (example 2: two factors, two levels → ‘**pills**’ or ‘**placebo**’ X ‘**sleep**’ or ‘**no sleep**’) | **Two-Way ANOVA**

ANOVA: the principles

- Like other classical statistical tests, we calculate a test statistic (the F-ratio) with which we can obtain **the probability of obtaining the data assuming the null hypothesis** (the P-value).

- A significant P-value (usually taken as $P<0.05$) suggests that **at least one group mean is significantly different from the others.**

Null hypothesis: all population means are equal

Alternative hypothesis: <u>at least one</u> population mean is different from the rest.

ANOVA: the principles

- ANOVA separates the variation in the dataset into 2 parts: **between-group** and **within-group**.

(NB: These variations are called the **sums of squares)**

- The **F-ratio** is then calculated as:

$$\frac{\text{Mean between-group SS}}{\text{Mean within-group SS}}$$

→ If the average difference between groups is **similar to that within groups**, **the F ratio is about 1**.

→ As the average difference between groups becomes **greater than that within groups**, **the F ratio becomes larger than 1.**

Linear models

- In order to perform an analysis of variance (ANOVA), you first need to **fit a linear model to your data.**

- Basically, your data becomes presented as a **formula** that resembles a mathematical function:

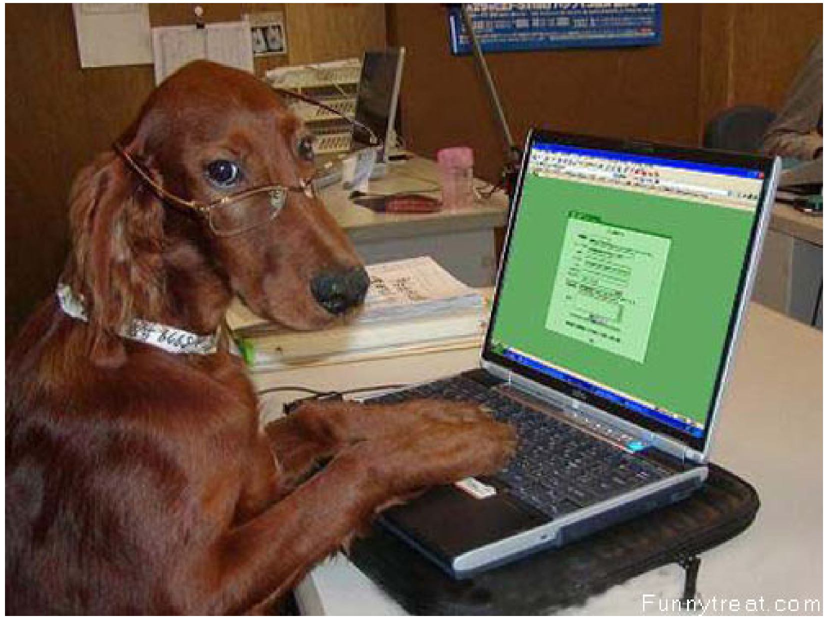
$$\mathbf{Y = B.X + U}$$

- Y is a matrix containing a set of measurements on each of the dependent variables

- X is your design matrix (observations on each of the independent variables)

- B is a matrix containing parameters (to be estimated)

- U is the error matrix, containing noise (**uncorrelated**, **normally distributed**)



Please progress forward to Step 3 !

Post-hoc tests

- The information that you get from an ANOVA is often incomplete: *at least one group mean is significantly different from the others* !

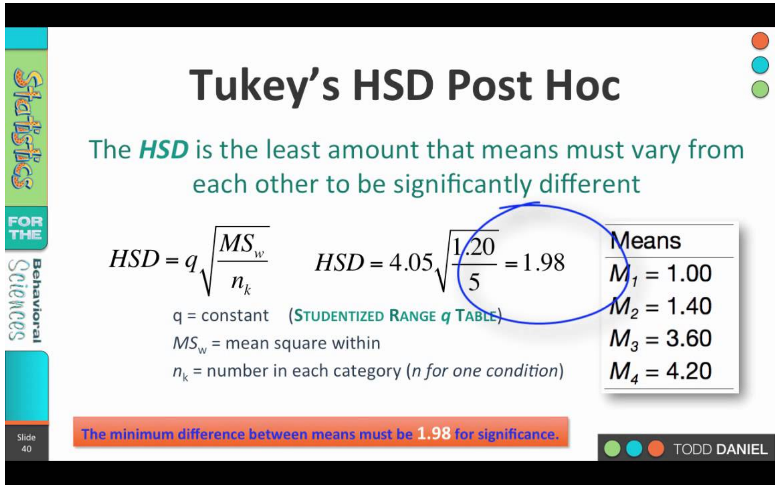
- For instance, you could be working with a factor that has **more than two levels** (Treatments: T1 to T4) and 'treatment' comes out as significant in your ANOVA. In that case, **you still don't know if all treatments are different from one another**, or if maybe only one is bad for the health, while the other three are comparable !

- In that case, you want to follow your ANOVA by **post-hoc tests**, which occur 'after' your main analysis.

Post-hoc tests

- One common and popular method of post-hoc analysis is **Tukey's HSD test** ('Honestly Significant Difference'). Tukey's test **compares the means of all treatments to the mean of every other treatment.**

- In general, HSD is preferred when you want to make all the possible comparisons between a large set of means (**six or more means**) and is considered the best available method when confidence intervals are desired, or **if sample sizes are unequal**.

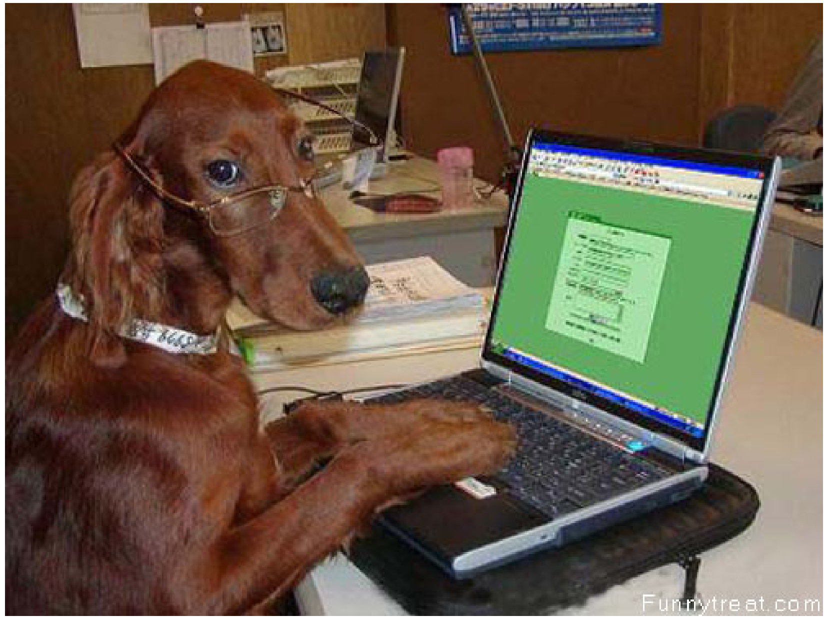


Post-hoc tests

- There are other ways to perform post-hoc tests, all of which should be able to be performed in R (either with base functions or via specific packages). For instance:

Fisher's LSD (Least Significant Difference): This test is the most liberal of all Post Hoc tests. The critical t for significance is unaffected by the number of groups. This test is generally not considered appropriate if you have more than 3 means.

Dunn's t-test: In general, this test should be used when the number of comparisons you are making exceeds the number of degrees of freedom you have between groups (e.g. K-1). This test is extremely conservative and rapidly reduces power as the number of comparisons being made increase.



Please progress forward to Step 6 !

Generalized linear models

- The ‘simple’ linear model that we’ve been using so far is actually a specific case of a broader class of linear models, which are called ‘**generalized**’ (because they are not limited to normally distributed data, and can be used in many other cases).

- In ‘generalized’ linear models (or GLMs), each outcome (Y) of the dependent variables is assumed to be generated from a particular distribution in an exponential **family**, that includes the normal (gaussian), as well as non-normal probability distributions.

- GLMs always contain a ‘**link function**’, which provides the relationship between the linear predictor and the mean of the distribution function.

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\beta = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$	$\mu = \mathbf{X}\beta$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$	$\mu = \exp(\mathbf{X}\beta)$

Generalized linear models

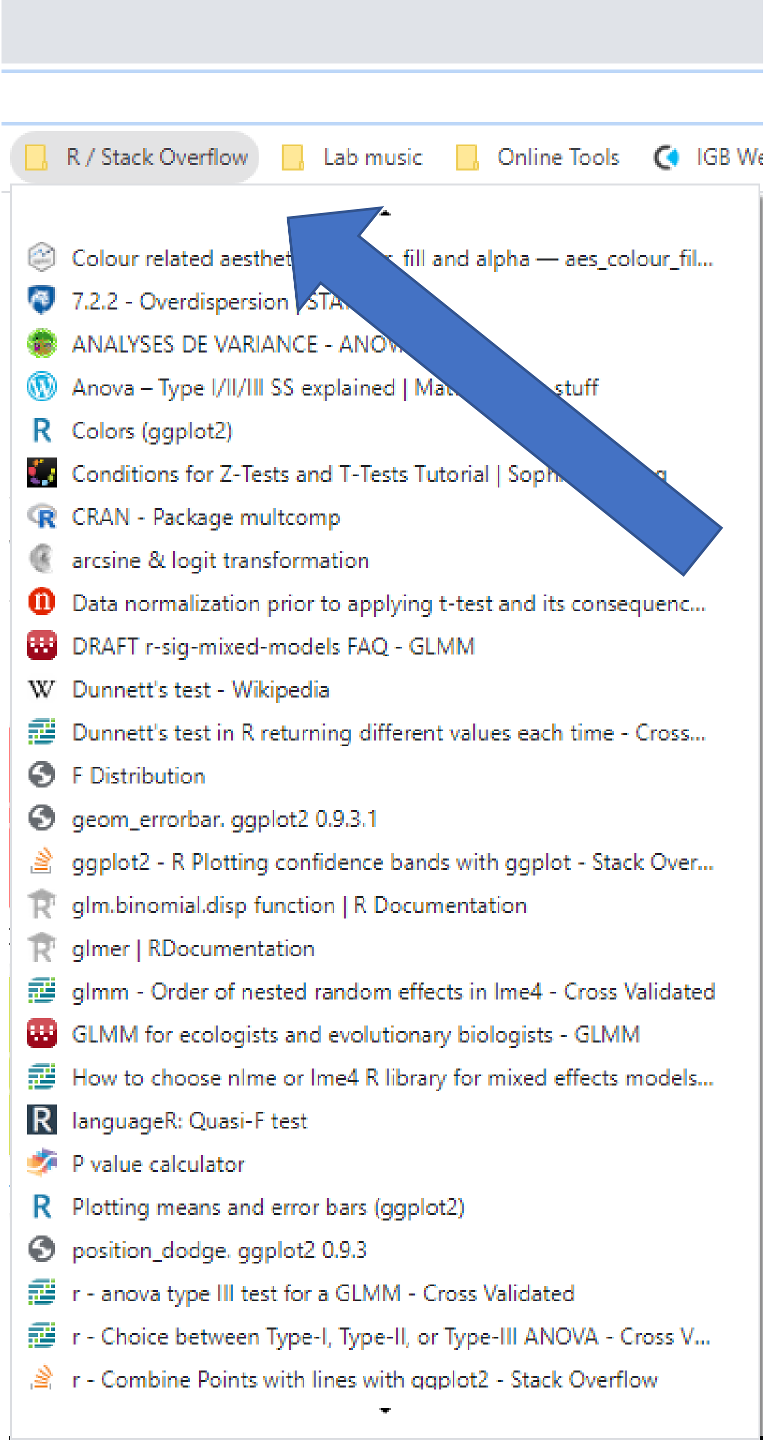
Generalized linear models are fit using the **glm( )** function. The form of the **glm** function is

glm(*formula***, family=***family***(link=***linkfunction***), data=)**

Family	Default Link Function
gaussian	(link = "identity")
binomial	(link = "logit")
Gamma	(link = "inverse")
poisson	(link = "log")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

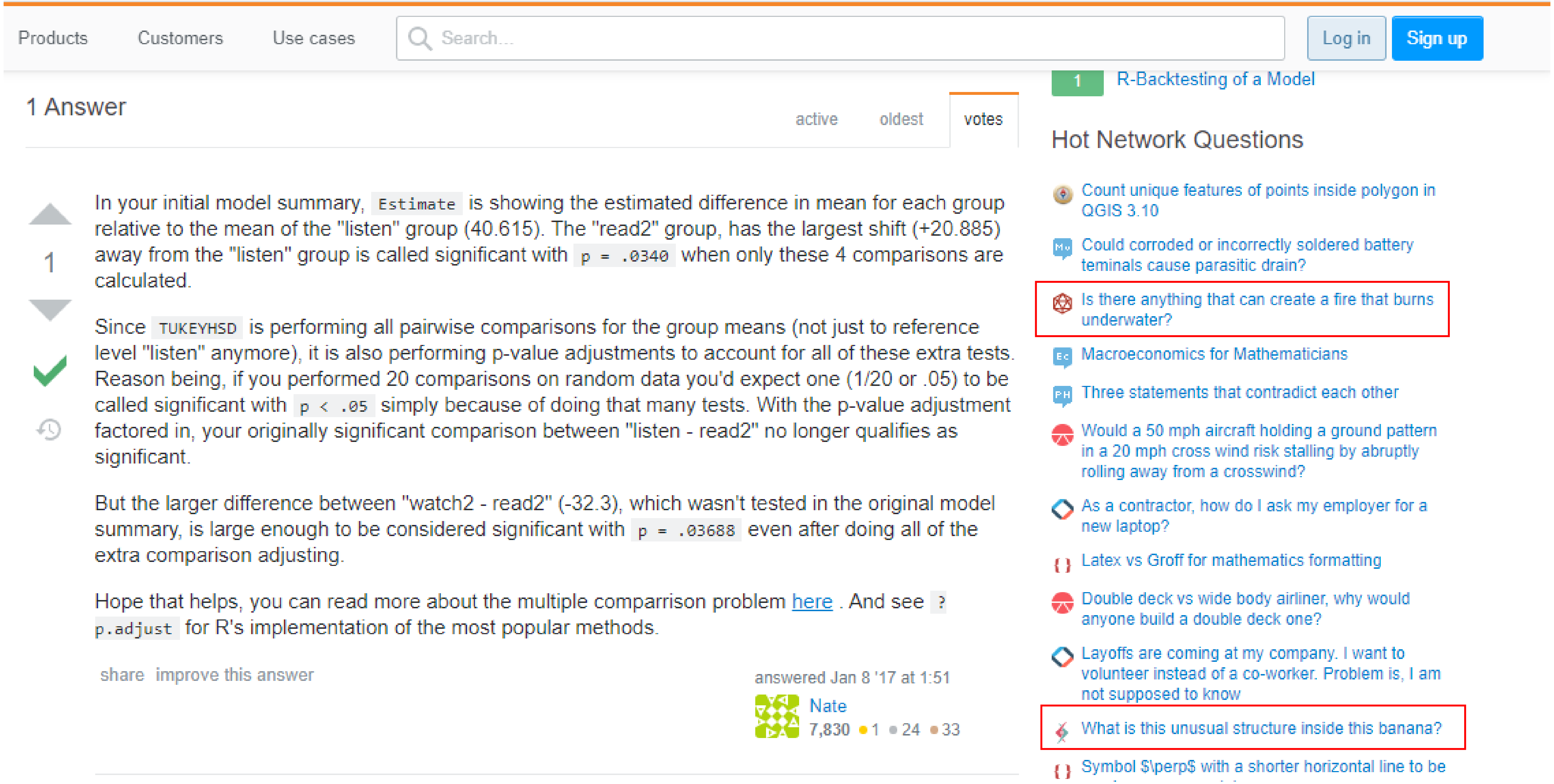
Useful tips:

- Stack Overflow is your friend !
- If you found a page helpful, **save it** under a safe directory !



This way, you won‘t have to look for it ever again !

Stack Overflow & distraction !



Products Customers Use cases Search... Log in Sign up

1 R-Backtesting of a Model

1 Answer

active oldest votes

1

In your initial model summary, `Estimate` is showing the estimated difference in mean for each group relative to the mean of the "listen" group (40.615). The "read2" group, has the largest shift (+20.885) away from the "listen" group is called significant with `p = .0340` when only these 4 comparisons are calculated.

Since `TUKEYHSD` is performing all pairwise comparisons for the group means (not just to reference level "listen" anymore), it is also performing p-value adjustments to account for all of these extra tests. Reason being, if you performed 20 comparisons on random data you'd expect one (1/20 or .05) to be called significant with `p < .05` simply because of doing that many tests. With the p-value adjustment factored in, your originally significant comparison between "listen - read2" no longer qualifies as significant.

But the larger difference between "watch2 - read2" (-32.3), which wasn't tested in the original model summary, is large enough to be considered significant with `p = .03688` even after doing all of the extra comparison adjusting.

Hope that helps, you can read more about the multiple comparrison problem here . And see `?p.adjust` for R's implementation of the most popular methods.

share improve this answer

answered Jan 8 '17 at 1:51

Nate

7,830 ●1 ●24 ●33

Hot Network Questions

- Count unique features of points inside polygon in QGIS 3.10
- Could corroded or incorrectly soldered battery teminals cause parasitic drain?
- Is there anything that can create a fire that burns underwater?
- Macroeconomics for Mathematicians
- Three statements that contradict each other
- Would a 50 mph aircraft holding a ground pattern in a 20 mph cross wind risk stalling by abruptly rolling away from a crosswind?
- As a contractor, how do I ask my employer for a new laptop?
- Latex vs Groff for mathematics formatting
- Double deck vs wide body airliner, why would anyone build a double deck one?
- Layoffs are coming at my company. I want to volunteer instead of a co-worker. Problem is, I am not supposed to know
- What is this unusual structure inside this banana?
- Symbol $\perp$ with a shorter horizontal line to be