ChatGPT and statistics

E-mail distributed on 11-11-2024

Dear all,

The use of ChatGPT has become widespread in science, including for answering statistics questions and producing R code. While its performance in these areas has admittedly become impressive, today I want to issue some cautions on that specific application, with first an important reminder of what ChatGPT actually is (and not is).

Large language models: Stochastic parrots

ChatGPT is an example of so-called "generative AI", alongside image generators such as DALL-E and Midjourney, and more specifically it is a "large-language model" (LLM). These models are not actually intelligent, in the sense that they cannot reason logically nor retrieve information. This may seem surprising considering that it seems to do exactly these things but it is important to keep in mind ChatGPT is merely a **predictive text generator**. Given a prompt, it will generate the text that is most probable to be its answer, **(a)** without understanding its content, and **(b)** regardless of its accuracy or veracity. This is why such models are sometimes called "stochastic parrots". Like a parrot, they are trained to give a response to a prompt, but merely as a surface imitation of human language. Where we see words, sentences, and communication, the bot sees only strings of numbers that maximize probabilities. Even the idea that we are giving ChatGPT "commands" is ultimately an illusion, although it is a very sophisticated one.

ChatGPT's inability to understand its own content is at the root of its problems with accuracy and veracity. By now the bot has become infamous for its so-called "hallucinations", answers in which it invents non-existent people, places, code, and scientific papers. While more recent versions have improved on this, the feature is unfortunately intrinsic to generative AI, and can never be fully regulated. The fact that ChatGPT hallucinates with the same apparent confidence as it gives factual answers is deeply concerning for many applications, and may be a reason why LLMs will turn out to be a passing phase in AI development.

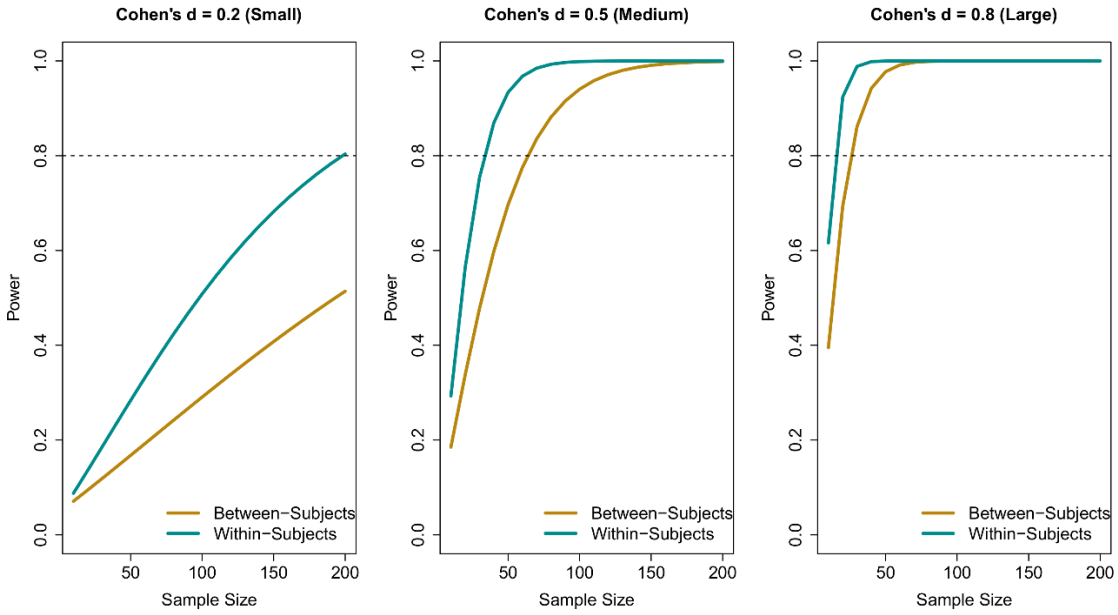
Statistics performance

Bearing the preceding caution in mind, ChatGPT currently performs quite well on statistics knowledge and R code, even for methods that are relatively obscure. I admit that I use it regularly at this point. For R, I find it especially useful for data manipulation and visualization. ChatGPT can generate

cumbersome scripts or shortcut functions that might take me half a day to figure out (badly). The bot is also good at producing demonstrations. For example, I entered the following prompt:

> How could I visualize a power curve in R for a simple two-group difference, contrasting a between-subjects design and a within-subjects design, for increasing effect size and sample size?

Which produced R code with explanatory comments that rendered this graph:



In other words, one can get educational content as well as direct analysis scripts. Especially if you are not an expert this is vastly more efficient than coding by yourself, or looking up the same answer on archives like StackExchange. Data manipulation and visualization is also a relatively safe application, in that code errors or incoherent output will be visible instantly. More tricky are direct theoretical questions, or code for complicated analyses. I quizzed ChatGPT on a couple of questions and got correct answers on all these problems:

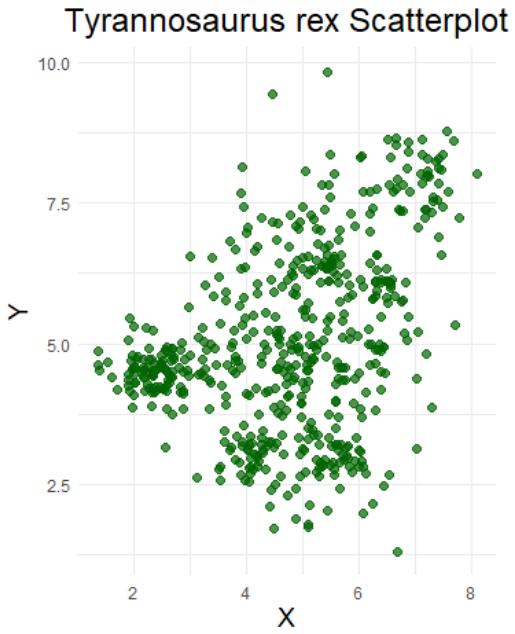
- What is the relation between the random intercept in a multilevel regression and the intra-class coefficient (ICC)?
- How to calculate BIC for a multivariate regression model?
- How to deal with structural zeroes in chi-square analysis?
- How to analyze square frequency tables for paired categorical data?

The BIC answer was surprising to me because ChatGPT returned the correct formula even though this is relatively obscure information. Nevertheless, the bot does also make mistakes.

Errors: Visible and invisible

Errors in ChatGPT come in a variety of flavors. As mentioned, erroneous R code is caught easily if the code simply does not run. The downside is that you may not have the expertise to properly debug it. A more worrying problem is R code that runs while being factually incorrect. Without the proper expertise, you would never realize that the output makes no sense.

For example, I asked ChatGPT how to calculate a *p*-value from a bootstrapped *t*-test, and it gave R code based on the logic of a permutation test, which is incorrect. In fact, when pointed out, ChatGPT immediately admitted the mistake and provided alternative code. Next, I asked ChatGPT to generate R code for a scatterplot where the points make the outline of a tyrannosaurus rex. The resulting code using the ggplot package produced this:



In addition, ChatGPT sometimes gives theoretically incorrect answers. In an older version, the bot misrepresented the ability of multilevel models to take into account serial correlation in longitudinal data, which it cannot do unless the user directly specifies, e.g., an autoregressive structure for residual correlation. I therefore urge caution to not accept ChatGPT's answers and R code without question, and to verify with an expert if you are working on a complex subject.

Hallucinations

Finally, there are the infamous hallucinations. I asked ChatGPT how to extract a model matrix from a structural equation model, and while it cited the correct package, "lavaan", it invented a non-existent "lavExtract" function to perform this operation. Next, I asked the bot to provide R code to run "agglomerative neighbor scaling", a method that does not exist. ChatGPT pretended that it did exist, although claiming that it was "less popular in R packages", and then proceeded instead to explain the similarly named methods of "agglomerative nesting" and "multidimensional scaling". When prompted for literature references, I received a mix of real and invented papers, including this one:

Okada, A., & Imaizumi, T. (2007). Multidimensional scaling of asymmetric proximity matrices with the parameterized factor analysis model. *Computational Statistics & Data Analysis*, *52(2)*, 837–851

While these authors work in this domain, no such paper was ever published. Older versions of ChatGPT would even invent entire journals and authors wholesale, in long lists of bogus citations. Such mistakes would be quickly caught when looking up the paper to read, but may be more dangerous when researchers carelessly copy-paste citations in their writing without checking them.

Conclusion

In conclusion, ChatGPT has a real practical use when it comes to answering statistics questions and generating correct R code. Therefore I do recommend it but with the cautions attached from this newsletter. Always consult with an expert if you are unsure of the bot's output, and always be aware of the inherent limitations of large language models.

Best,
Ben

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79