

Weekly Tip for Mar. 30, 2021

Statistics in R – Endless Possibilities Means Extra Caution

As a new R programmer, I was immediately struck by the stark differences in programming processes and techniques between SAS and R. Almost all of the more powerful procedures, and many basic ones, needed to be installed and called before you could use their features. I ran into very few instances of this in SAS, and when I did, the add-ons I included required only one installation and were forever accessible on that computer as long as I didn't move any folders around or uninstalled the base program.

Although sometimes frustrating, this ability to utilize multiple packages is still what I love the most about R. Being able to easily install and implement a package outside of the base program means that I can utilize statistical theory as it is developed! R packages are written not only by a large number of veteran data scientists, trying to find the most effective and efficient way of solving a problem, but also by theoretical statisticians who aim to make their latest research accessible to the scientific community through easy to implement procedures. This is extremely appealing to someone like me, who lives and breathes statistical theory.

However, this does lead to another pitfall – it is VERY easy to use the wrong procedure. In your quest to find the R package that has the easiest to implement procedure, you could end up applying the wrong theory and assumptions to your data! It is imperative that you double-check the mathematics underlying the process you are considering to make sure it matches your data. Unlike SAS, you not only have to test the assumptions of the basic models, but also the assumptions of your procedures. These are not always as well documented, so extra caution is needed.

Bottom line is, do not let your “efficient programmer” eyes be bigger than your data's stomach! When standing at the buffet of R packages, make sure that the ones you choose to put on your plate are the ones that are best for your data!

This week's tip was contributed by Deanna Schreiber-Gregory. Deanna is a government contractor and independent consultant who specializes in statistics, research methods, and data management.