RED THREAD®
by Ariadne Software

Comparing Cut Points Between Statistical Software Stata and R Using Random Effects Model

ariadnesolutions.com

The high-stake and high-risk nature of drug discovery and development demands either the use of a validated system or sufficient proof to support the use of an open-source system. A validated system warrants that the system is fit-for-purpose, and that it accurately, consistently, and repeatedly performs per the requirements and specifications established prior to development. Open-source software is available to the public at no or minimal cost for use, inspection, and modification. While commercial software may be validated, open-source systems are economical and preferable for ad hoc tasks. However, open-source software may or may not have been thoroughly tested and validated. While the FDA does not prohibit the use of open-source software, it does require reimplementation of the code to ensure the validity of the software output. Any modifications to either commercially available or open-source software, validated or not, must be further tested and validated.

Stata® and R are two commonly used software tools for performing statistical calculations in drug discovery and development. Stata is a statistical software package for data analysis, manipulation, management, and visualization. R is a programming language and environment for statistical analysis and visualization. While Stata is a validated software tool, R is open-source software tool with some test and validation scripts readily available. Not all R libraries are validated to the same extent as Stata®, and any modifications to either Stata or R require additional testing and validation. Regardless, in-house validation by the end-user is required for both, which is achieved by performing an Installation Qualification (IQ), Operational Qualification (OQ) and Performance Qualification (PQ) prior to use.

Purpose: The goal of this study was to confirm that the calculations made in the statistical software package R produced validated cut points (CP). Additionally, since the default settings for the commands utilized across statistical software packages can vary, comparing the calculations between two software packages helps ensure that the user's intentional choice of particular settings was suitable for the modeling used in CP determination. This comparison study used five datasets that originated from a variety of drug development stages – preclinical to clinical (three clinical study datasets and two nonclinical validation datasets). The datasets differed in not only the drug targets and moieties but also encompassed a variety of therapeutic treatments for diseases such as Triple Negative Breast Cancer, sickle cell anemia, along with others.

Study Number	Species	Study Type	No. Subjects	No of Panels/Iterations
Data 1	Preclinical	Validation	≤ 30	≤ 4
Data 2	Clinical	Validation	≤ 50	≤ 6
Data 3	Clinical	In-study	57	1 to 3
Data 4	Clinical	In-Study	53	1 to 3
Data 5	Clinical	In-Study	107	1 to 3

Table 1: Summary of the five datasets used for screening cut point calculations.

Approach: For the evaluation, a code was built in R using Linear Models for Linear Data (PLM) library and Shiny, an interactive server framework. The cut points were calculated using Random and Mixed Effects Model as follows:

General Random Effects Model (REM): $Y_i = \mu + U_i + W_i$

Where:
Y_i = S/N values
μ = average response of entire population
U_i = random effect model on the Gaussian distribution
W_i = individual-specific random error term

Mixed Effect Model (REM with Control for Fixed Effect (FE)): $Y_i = \mu + k + U_i + W_i$

Where:
k = fixed-effect terms for assignable parameter with k categories (e.g. number of panels) (known effect)

Results: By using a set of open-source packages for panel data methods, the cut point determination generated in Stata can be replicated utilizing the code in R. Comparable cut point values for a given distribution using the 95th and the 99th percentile in Stata and R were also observed. Building on that, post-modeling, R and Stata gave an equivalent new distribution at the 95th and 99th percentile. This equivalency was confirmed by plotting the CP values obtained from R vs Stata and performing a linear regression, which yielded a slope value of 0.9976 and a R2 value of 0.9998.

Cut Point Determination Approach	Study Number									
	Data 1		Data 2		Data 3		Data 4		Data 5	
	R	Stata	R	Stata	R	Stata	R	Stata	R	Stata
REM	5.25	5.26	1.48	1.47	1.40	1.47	1.49	1.50	1.33	1.32
MEM – Analyst FE	7.68	7.65	1.91	1.91	1.44	1.46	1.47	1.49	1.37	1.37
MEM – Iteration FE	4.75	4.75	1.51	1.50	1.40	1.44	1.49	1.50	1.33	1.33
MEM – Subject FE	2.51	2.51	7.33	7.33	1.21	1.19	1.39	1.36	1.37	1.37
MEM – Panel FE	4.02	4.00	1.80	1.87	1.42	1.40	1.42	1.40	1.28	1.29
MEM – Date FE	3.99	4.01	1.45	1.47	1.45	1.45	1.44	1.43	1.37	1.37

Table 2: A side-by-side comparison of cut points calculated using Stata and R for various combinations of Random Effects Model and Mixed Effect Models with various Fixed Effects.

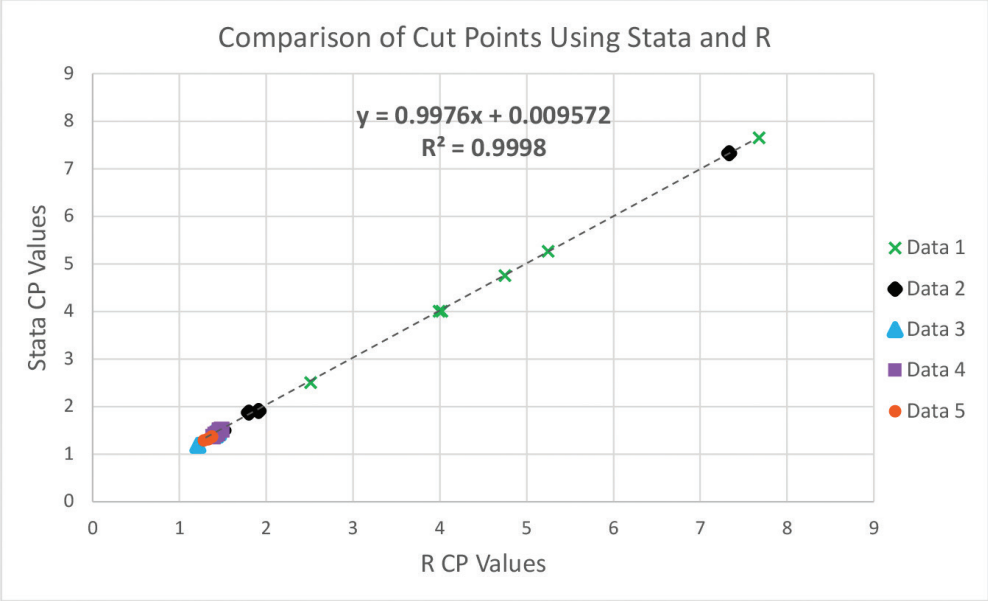

Figure 1: A scatter plot representing equivalency of cut points determined using Stata and R.

Conclusion: Using a set of open-source packages for panel data methods, an equivalent cut point determination can be made with the code in R than that made in Stata.

REFERENCES:

Schwartz, M., F. Harrell Jr, A. Rossini, and I. Francis. "R: Regulatory compliance and validation issues a guidance document for the use of R in regulated clinical trial environments." The R Foundation for Statistical Computing, c/o Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Augasse (2008): 2-6.

Stata is verifiably accurate https://www.stata.com/why-use-stata/verifiably-accurate/