Dear Robert and Didier,

I think there might be an issue for the non-paired test (e.g., one-way ANOVA and Kruskal test). I ran a simulation to illustrate why, please see below:

I used the actual MAPE data series by the base-lift method (average around 30%, with 100 observations), the data are not normally distributed. Then I generated an artificial data series by subtracting a small random number from each of the 100 observations. The subtracted number is between 0 to 1% and uniformly distributed. So these artificial data should have better performance than the base-lift method by definition.

Then I conduct the following tests:

1) One-way ANOVA,

2) Kruskal test

3) paired-t test

4) non-paired (independent)-t test

5) Wilcoxon SR test (paired)

The results:

All the non-paired tests fail to detect the difference.

One-way ANOVA and Kruskal test reported p-values of 0.51 and 0.34. The post-hoc test is not even needed because we have two series in this condition. Independent two-group t-test reported a p-value of 0.54, which failed to detect the difference as well.

However, Paired-t test and the Wilcoxon SR test both report p-values close to 0 (p-value < 2.2e-16), which rejected the null of no difference.

I also tried with completely simulated data which are normally distributed and found similar results. So I think it may not be the normality assumption which matters but the paired and non-paired assumption makes a big difference?

The Wilcoxon Sign Rank tests results for the models in our paper are shown below: most of the results are significant for IC, though moderate for EWC. However, I wonder do we include them or not (due to the limitation of applying statistical significance itself where were address by previous studies)?

Thank you very much!

Best regards,

Tao

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model/Forecast horizon/error measure | Forecast horizon= 1 | | | Forecast horizon= 4 | | | Forecast horizon= 12 | | |
| MAPE | SMAPE | MASE | MAPE | SMAPE | MASE | MAPE | SMAPE | MASE |
| ADL-own | ADL-own | ADL-own | ADL-own | ADL-own | ADL-own | ADL-own | ADL-own | ADL-own |
| ADL-own-IC | 0.015 | 0.133 | 0.273 | 0.000 | 0.019 | 0.079 | 0.001 | 0.001 | 0.003 |
| ADL-own-EWC | 0.949 | 0.246 | 0.114 | 0.494 | 0.680 | 0.498 | 0.365 | 0.392 | 0.579 |
|  | ADL | ADL | ADL | ADL | ADL | ADL | ADL | ADL | ADL |
| ADL-IC | 0.046 | 0.201 | 0.398 | 0.000 | 0.021 | 0.136 | 0.017 | 0.025 | 0.078 |
| ADL-EWC | 0.277 | 0.041 | 0.009 | 0.743 | 0.436 | 0.227 | 0.914 | 0.608 | 0.652 |
|  | ADL-DI | ADL-DI | ADL-DI | ADL-DI | ADL-DI | ADL-DI | ADL-DI | ADL-DI | ADL-DI |
| ADL-DI-IC | 0.081 | 0.530 | 0.287 | 0.060 | 0.333 | 0.243 | 0.147 | 0.303 | 0.192 |
| ADL-DI-EWC | 0.452 | 0.408 | 0.158 | 0.942 | 0.922 | 0.662 | 0.594 | 0.911 | 0.718 |