Validation of Visual Statistical Inference, Applied to Linear Models - Summary

Aarathy Babu

Statistical graphics are used in many places where numerical summaries simply do not suffice: model checking, diagnosis, and in the communication of findings. In this article, the lineup protocol
(the lineup is examined in the context of a linear model setting, which determines the importance of including a variable in the model), is compared with equivalent conventional test.

Difference between conventional test and visual inference

- Both methods start from the same set of hypotheses.
- The conventional test statistic is the t-statistic where as in the line up protocol has a plot of the data.
- In conventional hypothesis testing, the test statistic is compared with all possible values of the sampling distribution, the distribution of the statistic if the null hypothesis is true. If it is extreme on this scale then the null hypothesis is rejected.

In visual inference, the plot of the data is compared with a set of plots of samples drawn from the null distribution. If the actual data plot is selected as the most different, then null hypothesis is rejected.

- lineups will depend on observers' evaluation.

Assessing an individual's skill to identify the actual data plot will require that an individual evaluate multiple lineups.

How is p value of a line up calculated ?

Under the null hypothesis, each observer has a 1/m chance of picking the test statistic from the lineup.

- K = independent observers
- X = number of observers picking the test statistic from the lineup
- null hypothesis =
$$X = Binom_{K,1/m}$$

$$P(X \geq x) = 1 - Binom_{K,1/m}(x-1)$$

- x = number of observers selecting the actual data plot

Power

Probability to reject the null hypothesis for a given parameter value

$$Power_V(\theta) = Pr(Reject H_0|\theta)$$

For two different visual test statistics of the same actual data, one is considered to be better, if the actual plot is more easily distinguishable to the observer. Power is typically used to measure this characteristic of a test.

How does observer skills and size of line up matter ?

A mixed effects logistic regression model , accommodating both for different abilities of observers as well as differences in the difficulty of lineups are used.

for larger size m, the probability of correctly identifying the actual data plot decreases with m.

Experiments with simulated data

Three experiments are conducted :

The first two experiments have ideal scenarios for conventional testing, where we would not expect the lineup protocol to do better than the conventional test. The third experiment is a scenario where assumptions required for the conventional test are violated, and we would expect the lineup protocol to outperform the conventional test.

Data cleaning done by using a test plot (easy plot) which everyone should get correct, so that a measure of the quality of the subjects effort could be made.

Experiment 1 : Discrete Covariate

- To study the ability of human subjects to detect the effect of a single categorical variable.

Result : - visual inference mirrors the power versus effect relationship of conventional testing.

- In Experiment 1, there is more variability between subjects, with some doing better than the conventional test on large effects.

Experiment 2 : Continuous Covariate

This experiment is very similar to the previous one, except that there is a single continuous covariate and no second covariate.

Result : - visual inference mirrors the power versus effect relationship of conventional testing.

- In Experiment 2, subjects performed similarly, and substantially better than the conventional test

As the p-value increases the proportion of correct responses falls in both expt 1 and 2, shows direct association between proportion of correct responses and conventional test p-values.

people tended to select the plot with the lowest p-value.

Experiment 3 : Contaminated Data

The simulation is conducted in the hope that the visual test procedure, will at least compare favorably with the conventional test—without any ambition of performing equally well.

Result:

- the power of the visual test exceeds that for the conventional test, as expected.
- In Experiment 3, there is the most subject-specific variation. Some subjects performed substantially better than the conventional test, and on average the visual test was better.
- visual p values are very small no matter what the conventional p-values are. conventional test loses its power to reject H0 even when the alternative is true, whereas the visual test performs well.

How Much Do Null Plots Affect the Choice?

- If the actual data plot is very different from all of the null plots, then the null plots should not have much influence on the choice.
- If there is a null plot with a small p-value, or one close to that of the actual data plot, we would expect that subjects have a harder time detecting the actual data plot.

Type III Error

- Type III errors are defined as the probability of correctly rejecting the null hypothesis but for the wrong reason.

Conclusions

- The power of a visual test increases with the number of observers.
- the theoretical power of visual test can be better than that of conventional tests.