EAI Endorsed Transactions
on Context-Aware Systems and Applications

Research Article EAI.EU

A Framework for Utilizing Permutational Multiple Analysis of Variance as a Precursor for Nonparametric Statistical Learning with Cyber Network Data

T.A. Woolman[1,*] and J.L Pickard[2]

[1]1st Indiana State University, Terre Haute, IN, USA
[2]2nd East Carolina University, Greenville, NC, USA

Abstract

INTRODUCTION: Although scientific hypothesis testing methodologies are well established, their application to falsifiable hypothesis testing for assessing causal relationships potentially identified by machine learning and artificial intelligence models is rare due to the primarily nonparametric statistical nature of these systems.

OBJECTIVES: The primary objective of this study is to demonstrate the potential for applying nonparametric statistical tests to a mixed qualitative and quantitative cyber network dataset as a method to pre-assess the feasibility of applying forms of statistical hypothesis testing before a machine learning algorithm models the data.

METHODS: A mixture of permuted analysis of variance models augmented by the use of transformed non-Euclidean multivariate distances between curated dependent variable classes produced this research data. Quasi-experimental data from an enclosed laboratory environment utilizing a monitored, locally unrestricted network that introduced known Internet of Things (IoT) malware software supplied network flow events.

RESULTS: A PERMANOVA model was executed against 62,000 records of the network flow observations, using Euclidean distance measurements with variable-dependent relationship ordering, using terms added sequentially (first to last) in the order encountered in the raw network flow dataset, using 200 permutations. This precursor test resulted in a p-value for the PERMANOVA model that incorporated terms added sequentially of 0.02985, providing an F value of 0.00017 with which to determine the ratio of explained to unexplained variance. Utilizing an analysis of the F values for all of the residuals, we show 29,998 degrees of freedom with a residual F model score of 0.99983, indicating that there is a strong proportion of explained to unexplained variance across all of the independent variables contained in the model. The model is thus statistically significant with a p-value below the alpha test statistic of 0.05.

CONCLUSION: This research has demonstrated that it is possible to apply tests of falsifiability that incorporate reproducible methods into the quasi-experiment design and apply this to the field of machine learning. Applied to AI/ML (artificial intelligence/machine learning) models, this pre-assessment methodology supports the appropriateness of cyber network flow datasets in which a final test of statistical significance would be required. The authors believe that this represents a substantially useful precursor assessment stage for the suitability and reliability of the utilization of any nonparametric statistical learning algorithms applied to cyber network data predictive analytics.

[*] Corresponding author. Email: twoolman@ontargettek.com

Keywords: Artificial intelligence, machine learning, Internet of Things, malware, network flows, multinomial classification, falsification, PERMANOVA, asymptotic, permuted, NPMANOVA

Received on 12 December 2022, accepted on 19 May 2023, published on 12 July 2023

doi: 10.4108/eetcasa.v9i1.2929

1. Introduction

The purpose of this study is to determine the potential suitability for the future development of a nonparametric statistical learning predictive classification models utilizing deep learning and machine learning algorithms to investigate issues related to the detection of malware and network intrusion attacks against internet of things (IoT) devices on unrestrained network connections

To address the cyber defence shortcomings of the hardware devices comprising the IoT we propose the use of a permuted multiple analysis of variance (PERMANOVA) as a nonparametric test of statistical significance as a “pathfinder” or precursor model prior to the implementation of a machine learning or deep learning predictive classification model for network flow data.

There is an overall lack of research in the extant literature regarding the application of the scientific method in terms of the use of falsifiability hypothesis testing with which to determine the statistical significance for the effective application of deep learning and machine learning algorithms for this and related use cases involving network flow datasets. This is due to a fundamental conflict between the nature of classic and inflexible linear regression models with which to conduct statistical hypothesis testing and nonparametric deep learning models.

This research utilizes nonparametric tests of statistical significance utilizing permutational analysis of variance (PERMANOVA) to determine if linear and nonlinear relationships exist between a suite of independent variables and a single, curated labelled dependent multinomial malware class variable in a large network flow dataset obtained from a series of unrestricted network environments.

Results of this study show that the correct implementation of a PERMANOVA as a potential AI/ML pathfinder model can become an intrinsic component of a scientific falsifiability test for mixed qualitative and quantitative complex cyber datasets, and potentially lead to cost savings through the economic and appropriate utilization of AI/ML model development expenses.

1.1. Background

As the Internet of Things is projected to reach approximately 27 billion active connections by 2025 [1] (IoT.Business.News, 2022), a growing imperative exists to help secure this increasingly critical aspect of the global digital infrastructure. Traditional rules-based and heuristic-based cyber defensive systems are often inadequate to address the unique challenges, risks, and limitations inherent in IoT communications architectures. Consequently, the need for sophisticated, adaptable, and defensive cyber systems as part of a defence in depth strategy that is scientifically proven to be demonstrably effective across a broad range of risk categories utilizing falsifiable hypothesis testing has never been greater.

The IoT-23 dataset, also known as the Avast IoT-23, used for this study is a curated human-labelled IoT dataset produced by Stratosphere Laboratory, AIC group, FEL, CTU University, Czech Republic. It utilizes human-labelled curated analysis of individual network traffic flows from IoT devices in controlled environments consisting of 20 malware captures and three captures for benign IoT devices traffic. The dataset produces labelled flow data for known, authenticated Internet of Things malware infection activities, as well as benign network behaviours recorded on unrestrained network environments for detailed analysis.

1.2. Problem statement

The problem addressed by this study is the need for sophisticated, adaptable defensive cyber systems as part of a defence in depth strategy to address the limitations of traditional rules-based and heuristic-based cyber defensive systems which are often inadequate to address the unique challenges, risks, and limitations inherent in IoT communications architectures.

Previous research on the security design and unique aspects of Internet of Things (IoT) devices has indicated an ongoing challenge in detecting and classifying a growing range of malware threats that employ any falsifiable, scientific hypothesis testing in their methods that utilize nonparametric statistical learning, including various types of deep learning predictive models. Recent studies have indicated that there are several static-based methods for IoT malware detection, including analysis of opcode features, string data, ELF headers from binary data files, as well as other methods related to the analysis of binary executable machine language code and function calls.

There is currently no generally accepted mechanism with which to demonstrate scientifically with permuted statistical certainty that a nonparametric statistical learning algorithm is applicable for addressing a null hypothesis statement related to an IoT network flow dataset. Therefore, the objective of this research is to propose a method to address the problem of testing for permuted statistical significance in a dataset. Specifically for a dataset in question containing IoT network flow traffic supporting a null hypothesis test related to AI/ML malware classification.

In doing so, the aim of this research is to propose a benchmark scientific falsification process that can be utilized in a pre-test environment, prior to the undertaking of implementing deep learning or machine learning predictive malware classification models in IoT network flow data.

1.3. Research questions and hypothesis

Hybrid linear and nonlinear statistical analysis is used to provide a framework to ascertain the statistical significance of the relationships present in complex, multivariate network flow traffic data. A Permuted variant of traditional statistical hypothesis testing provided a measure of falsifiability of the relationship between the independent variables having a causal effect on the dependent variable classes. The intent of this test is to act as a precursor method, prior to the employment of more complex predictive machine learning and deep learning models. The intent is to provide supporting evidence for the presence of statistically significant causal relationships to be present in a random selection of the overall network traffic dataset and to ascertain the potential falsifiability of a nonparametric statistical learning algorithm before its application to training on this dataset. The substantive research questions area:

RQ1: What do the network flow variables reveal about the distribution of malware classifications in the IoT-23 dataset?

RQ2: What is the ability to construct nonparametric predictions for the known malware classes?

The objective for both findings is to scientifically support the suitability of the data supporting the appropriateness of the use of supervised machine learning and deep learning predictive methodologies, including multinomial classification predictive models for malware detection using network flow data in this use case.

Answering these research questions requires the construction of a nonparametric permuted analysis of variance (NPMANOVA), also known as a permuted analysis of variance

(PERMANOVA) model. The PERMANOVA model will be used to test the following hypotheses:

H10: The centroids, or the vector of means, of each of the distinct malware groups' multivariate dependent network flow variables in the IoT-23 dataset are equal.

H1a: There is at least one pair of malware groups with significantly unequal multivariate dependent variable centroids in the network flow dataset.

1.4. Significance of the study

Historically, hypothesis testing as a component of research that utilized machine learning and deep learning prediction models has largely been absent because the nonparametric nature of these algorithms was often in conflict with the linear relationships that were required between dependent and independent variables for the development of valid inferences for hypothesis testing. Recent developments in both explainable AI technology frameworks as well as advances in the use of subsampling techniques and variance estimation, under suitable conditions, can potentially allow for the creation of asymptotically normal predictions that can meet the criteria for some forms of hypothesis testing. The proposed precursor model framework will address this falsifiability gap as it applies to the research question hypotheses.

Because of the complex nature of IoT network traffic, coupled with the increasing number of malware attack scenarios and the multi-architecture nature of many IoT devices, a nonparametric, permuted statistical analysis technique will create a "pathfinder" test model. The purpose of this analytical model is to conduct a test of statistically significant causality between the suite of independent variables across the entire dataset and the multinomial factor dependent variable after it has been dummy encoded into an ordinal value.

The need for scientifically falsifiable as well as adaptable, robust, and scalable real-time malware detection and classification for Raspberry Pi IoT devices in unrestrained Internet-connected networks is apparent, especially given the tremendous growth and near-omnipresence of these devices in both industry, hospitals, and home network environments.

1.5. Assumptions

We intend to develop an ensemble of deep learning classification and prediction models based on various cutting-edge advanced neural networks for the purpose of classification and probabilistic scoring of known types of IoT malware network intrusion threats for the Raspberry Pi ARM-based processor hardware. The original datasets for the quasi-experiment that was conducted consisted of hundreds of gigabytes of unencrypted TCP and UDP internet protocol network traffic data, freely provided via the Stratosphere Laboratory [2], using the Aposemat IoT-23 labelled dataset with malicious and benign IoT network flow traffic. The primary assumption in this research is that any random sampling conducted from the IoT-23 network flow datasets to achieve a stated statistical power using a Chi-square Goodness of Fit Test are representative of the total population contained in the IoT-23 data.

1.6. Background

The primary limitation for this study is that it will be based upon the IoT-23 dataset, and thus be an ex-post facto experiment with relatively limited (but still substantial) network flow traffic data. While this dataset is robust and well curated, the characterization of findings obtained from the models produced from the relationships in this data may not adequately capture a comprehensively state of the art or fully global perspective on the most modern IoT malware threats.

One limitation of the use of PERMANOVA within the study is the memory-intensive nature of the distance measurements used between observations. [3] states that the analysis of variance employed utilizes distance matrixes to conduct a partitioning of the data to determine sources of variation taking place. Those distances are then used to fit linear models such as factor and polynomial regression to the distance parameters to produce permutation tests that ultimately provide pseudo-F ratio scores.

1.7. Previous studies

IoT devices are largely computationally resource constrained and have access to relatively little on-board memory while also contending with small power supplies due to constrained energy storage, diminutive power, distribution, and restrictive power scavenging capabilities [4]. Likewise, some of the challenges related to malware detection and classification for Internet of Things (IoT) devices using a network intrusion detection system (NIDS) is that IoT network traffic largely consists of homogenous protocols, hardware, and software components [5].

[6] Stated that the vast (and rapidly growing) number of IoT devices, inherently poor network security processes and their typical reliance on unrestrained permanent Internet connections have made IoT devices a convenient tool for threat actors to infect these devices and then organize them for powerful cyberattacks. [7] expanded on this by discussing how many IoT device manufacturers are focused on enhancing profitability by making these devices as inexpensive and quickly manufactured as possible and sacrificing many security elements in their design as a result.

According to [2], the IoT-23 dataset is a dataset based on curated, human-labelled network flow data that consists of internet protocol network traffic flows from Internet of Things (IoT) devices. These IoT devices rely on the Raspberry Pi hardware architecture standards as applied to a range of specific consumer electronic devices, connected to network host devices.

The primary type of IoT hardware platform utilized in this series of deep learning AI quasi experiments for data capture and malware validation was the Raspberry Pi IoT device type, utilizing extensive third-party datasets of Internet network traffic data [2] for both benign and permutations of malware infestations on these devices. This quasi-experiment exclusively used network traffic flow captures. This provided a means for utilizing inexpensively obtained low-storage requirement data that potentially contained sufficient attributes and variabilities for identifying potential malware events, while not relying on more costly packet capture storage and analysis systems.

2. Methodology

The research was conducted utilizing the R statistical programming language and the RStudio development environment [8], along with a set of open-source deep learning package library algorithms.

The potential utility of this proposed methodology of utilizing a precursor NPMANOVA/PERMANOVA test is to ascertain if dataset in question has nonparametric predictive utility that could then potentially be augmented further using nonparametric statistical learning applications.

Reproducibility for this research relies primarily on the utilization of the Vegan package for R as discussed in [3] that primarily feature peer-reviewed tools for nonparametric diversity analysis as well as ordination methods and methods for dissimilarity analysis. Specifically for this research, the experiment utilized permuted multivariate analysis with the adonis2 function within Vegan using distance matrices. It incorporated a permutation test, in this case using 200 permutations (as both a generally accepted allowable convention as well as a limitation due to the computational expense related to permuted tests against such a large dataset).

The adonis2 function was used in its capacity for the analysis of dissimilarities through partitioning of sums of squares as discussed in [9] for using distance-based measures for fitting permuted multivariate data models and is referred to by the package authors as "permutational MANOVA", formerly known as nonparametric MANOVA.

The permuted MANOVA then assess the statistical significance of a pseudo F-statistic, an F-ratio similar in nature used within an ANOVA study. This pseudo statistic relies on ranked dissimilarities of objects between groups (assuming similar measures were used between groups), to that of the variation of those same objects (again using similar measures) within the same group. In this manner, the nonparametric variation is comparable to both within and between groups using the same object measures. The interpretation of the F-ratio statistics indicates the amount of variation of means within and between group dispersions and provides a p-value measure of statistical significance.

A Chi-square Goodness of Fit Test determined that a randomly generated n=62,000 IoT-23 flow observations would provide a sufficient statistical power for small effect size measurement using the dependent variable categorical factors present. This random sample population allowed for the economization of data processed by the adonis2 function within the Vegan package due to available computing resources, while still enabling a sufficient statistical power for detecting small effect sizes with the number of dependent variable factor classes encountered. The process to produce a visualization of the dependent variable group dissimilarities began by first calculating the Bray-Curtis distances between groups using the "vegdist" (vegan distances) function in the Vegan package. The Bray-Curtis distances determined the multivariate dispersion between groups using the "betadisper" (beta dispersion) function supplied by Vegan to calculate multivariate dispersions across principal component axes. The results from "betadisper" provided the multivariate centroids for each group within the dependent variable in the random subsample dataset, producing the Homogeneity of Multivariate Dispersions table shown in the results. This table provided both positive and negative Eigenvalues across each principal component axis generated by the Beta Dispersion plot between groups using the non-Euclidian Bray-Curtis ("Bray") distances.

3. Results

An initial PERMANOVA model was executed against the 62,000 records of the IoT-23 network flow observations, using Euclidean distance measurements with variable-dependent relationship ordering, using terms added sequentially (first to last) in the order encountered in the raw IoT-23 network flow dataset, using 200 permutations. This initial test resulted in a p-value for the PERMANOVA model that incorporated terms added sequentially of 0.02985, providing an F value of 0.00017 with which to determine the ratio of explained to unexplained variance.

Utilizing an analysis of the F values for all of the residuals, we show 29,998 degrees of freedom with a residual F model score of 0.99983, indicating that there is a strong proportion of explained to unexplained variance across all of the independent variables contained in the model. The model is thus statistically significant with a p-value below the alpha test statistic of 0.05. We can thus reject the null hypothesis statement that, "The centroids of each of the distinct malware groups' multivariate dependent network flow variables in the IoT-23 dataset are equal." We can conclude that this dataset sample likely contains patterns that a nonparametric statistical learning algorithm can utilize with which to conduct successful AI/ML predictive modelling, allowing for model development with a higher confidence of scientific proof.

According to [10], the PERMANOVA methodology as encapsulated by the adonis function of the vegan package [3] employs a calculation of the transformed centroid of the sequentially combined independent variables, and then determines the squared deviation for each distinct group factor ("treatment group") within the dependent variable to that centroid. The methodology then conducts tests of statistical significance utilizing a pseudo F-test from the sequential sum of squares based don permutations of the available raw data sample set that were provided.

The potential benefit shown by this PERMANOVA precursor methodology include both demonstrating the applicability of nonparametric statistical learning ("machine learning") algorithms to this dataset within the defined use case. Likewise, the ancillary benefit of this methodology framework would be the ability to quantify the improvement gains from a self-optimized nonparametric statistical learning algorithm when compared to the NPMANOVA precursor test as a baseline comparison. Potential negatives of this methodology framework include additional analytical stages necessary as well as computational cost increases.

As a means of visualization of these centroid deviations for each treatment group (multi-level factor variable dependent variable), an analysis of Homogeneity of Multivariate Dispersions was conducted utilizing methods from [3], a permuted multivariate variation of Levene's test for homogeneity of variances. Due to the permuted multivariate nature of this test, the methodology is extremely computationally intensive and was beyond the capabilities of available computing resources. This phase of the experiment therefore required the use of randomized balanced class statistical subsampling from the larger original dataset to conduct the beta dispersion analysis. A statistically balanced sample size of n = 500 observations for each curated class group in the dependent variable was randomly selected and used for this analysis.

The result of this further permuted test demonstrated a significant multivariate dispersion that was measurable for the majority of malware and benign class groups within this subsampling, as shown in Figure 1.

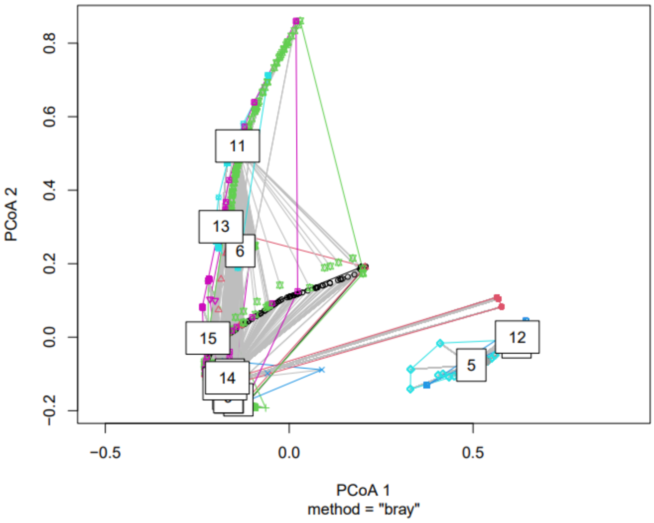

Figure 1. Permuted beta dispersion of human-curated balanced malware (DV) classes using PCoA transformed axes, indicating the significant non-homogeneity of malware events.

Figure 2 two shows the 2-dimensional principle coordinate analysis plot illustrating the transformed non-Euclidean multivariate distances between measurable malware class groups in the dependent variable, visually demonstrating compactness within and non-Euclidean separations between groups. The Bray-Curtis Dissimilarity Beta Diversity method measure variances for between-groups composition medians across multiple independent variables for a given sampling population, employed using randomized class balancing for unequal group sizes (n=500).

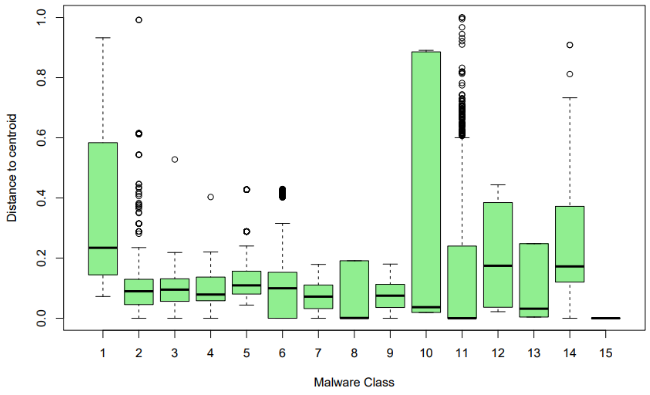

Figure 2. Boxplot of malware class distances to centroids.

4. Conclusion

This research has demonstrated that it is possible to apply tests of falsifiability that incorporate reproducible methods into the quasi-experiment design and apply this to the field of machine learning for the pre-selection of appropriate cyber network flow datasets that could be potential candidates for an AI/ML predictive model. Specifically, the authors successfully applied a permuted MANOVA/ANOVA technique to demonstrate statistical significance in the relationship of sequentially added network flow independent variable terms. This scientifically demonstrates against the stated alpha test statistic that a nonlinear relationship exists between the independent variables in the network flows and our dependent set of malware class groups that were contained within an encoded multi-level categorical factor dependent variable representing distinct treatment groups or malware classes, including the benign states. To further support statistical reliability, the number of observations included within the PERMANOVA for statistical significance testing utilized a Chi-square Goodness of Fit Test for determining *n*.

The positive impacts of this approach include a stronger assurance that the data contained in the independent (predictor) variables have established, non-random statistical patterns that are either linear or non-linear (or contain aspects of both) that are potentially detectable by a nonparametric statistical learning algorithm.

Potential negative impacts with this practice are the increased computational burden and analytical phases associated with incorporating the methodology into a predictive modelling architecture.

The authors believe that this represents a substantially useful precursor assessment stage for the suitability and reliability of the utilization of any nonparametric statistical learning algorithms applied to cyber network data predictive analytics. The authors propose this methodology as a framework to govern costs and development risks (success uncertainty) associated with enterprise-scale AI/ML model development efforts as it applies to cyber network data. This in turn will improve the likelihood of developing successful AI/ML systems that are capable of detecting malware within network flow observations.

5. Recommendations for Future Research

The novelty of utilizing permuted MANOVA/ANOVA methodologies and utilizing techniques such as dissimilarities using beta diversity methods to conduct nonparametric transformations of complex multivariate datasets represents an opportunity to demonstrate nonlinear tests of homogeneity in cyber network data. These tests produce a variety of explainable and readily visualized digital signatures for a wide range of cyber network threats.

The authors believe that this framework can not only be used to assess the potential applicability for accurate AI/ML supervised malware detection models but may also be used as a means of falsifiability determining statistical significance of these post-hoc experiment models.

Further studies involving permuted, nonparametric statistical methods should explore the potential suitability, accuracy, explainable findings and provability of nonparametric statistical learning algorithms. Specifically, we recommend exploring the further potential of permuted scientific research applications of AI/ML models for cyber intrusion detection to improve both utilization efficiency, reliability, and scalability of AI/ML model assurances.

References

[1] IoT.Business.News. State of IOT 2022: Number of connected IOT devices growing 18% to 14.4 billion globally [Internet]. IoT Business News. 2022 [cited 2022Dec6]. Available from: https://iotbusinessnews.com/2022/05/19/70343-state-of-iot-2022-number-of-connected-iot-devices-growing-18-to-14-4-billion-globally/

[2] Stratosphere Laboratory. (2020). Aposemat IoT-23. A labeled dataset with malicious and

benign IoT network traffic. Parmisano, A., Garcia, S., Erquiaga, M. J. Available online at https://www.stratosphereips.org/datasets-iot23. Accessed on October 17, 2020.

[3] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, Simpson GL, Solymos P, Stevens MH, Wagner H, Oksanen MJ. Package 'vegan'. Community ecology package, version. 2013 Dec 12;2(9):1-295.

[4] Raj A, Steingart D. Power sources for the internet of things. Journal of the Electrochemical Society. 2018 Apr 25;165(8):B3130.

[5] Anthi E, Williams L, Słowińska M, Theodorakopoulos G, Burnap P. A supervised intrusion detection system for smart home IoT devices. IEEE Internet of Things Journal. 2019 Jul 2;6(5):9042-53.

[6] Bobrovnikova K, Lysenko S, Gaj P, Martynyuk V, Denysiuk D. Technique for IoT Cyberattacks Detection Based on DNS Traffic Analysis. InIntelITSIS 2020 (pp. 208-218).

[7] Murphy M. The Internet of Things and the threat it poses to DNS. Network Security. 2017 Jul 1;2017(7):17-9.

[8] Chizinski C. Permutational multivariate analysis of variance using distance matrices (Adonis) [Internet]. Christopher Chizinski. 2014 [cited 2022Dec6]. Available from: https://chrischizinski.github.io/rstats/adonis/

[9] McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001 Jan;82(1):290-7.

[10] Chizinski C. Permutational multivariate analysis of variance using distance matrices (Adonis) [Internet]. Christopher Chizinski. 2014 [cited 2022Dec6]. Available from: https://chrischizinski.github.io/rstats/adonis/