MONASH University

Advances in Artificial Intelligence for Data Visualization: Developing Computer Vision Models to Automate Reading of Data Plots, with Application to Predictive Model Diagnostics

Weihao Li

B.Comm. (Hons), Monash University

A thesis submitted for the degree of Doctor of Philosophy at

Monash University in 2022

Department of Econometrics and Business Statistics

Contents

Copyright notice

(Choose one of the following notices.)

(Notice 1)

© Weihao Li (2022).

The second notice certifies the appropriate use of any third-party material in the thesis. Students choosing to deposit their thesis into the restricted access section of the repository are not required to complete Notice 2.

(Notice 2)

© Weihao Li (2022).

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Declaration

(Standard thesis)

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

(Thesis including published works declaration)

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes (insert number) original papers published in peer reviewed journals and (insert number) submitted publications. The core theme of the thesis is (insert theme). The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the (insert name of academic unit) under the supervision of (insert name of supervisor).

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.) Remove this paragraph for theses with sole-authored work

In the case of (insert chapter numbers) my contribution to the work involved the following:

Thesis chapter	Publication title	Status (published, in press, accepted or returned for revision)	Nature and % of student contribution	Co-author name(s), nature and % of co-author's contribution	Co-author(s), Monash student Y/N
2	xx	xx	xx	xx	N
3	xx	xx	xx	xx	N
4	xx	xx	xx	xx	N
5	xx	xx	xx	xx	N

I have / have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: Weihao Li

Student signature:

Date:

Acknowledgements

I would like to thank my pet goldfish for …

Preface

The material in Chapter 1 has been submitted to the journal *Journal of Impossible Results* for possible publication.

The contribution in Chapter **??** of this thesis was presented in the International Symposium on Nonsense held in Dublin, Ireland, in July 2015.

Chapter 1

Introduction

Chapter 2

Automatic Visual Statistical Inference, with Application to Linear Regression Diagnostics

2.1 Abstract

2.2 Introduction

2.2.1 Model Diagnostics

Model diagnostics is the part of data analysis, preceded by the fit of a model, whose primary objectives are to examine the goodness of fit and reveal potential violations of assumptions. In these diagnostics, though numeric summaries are available and some are endorsed by finite or asymptotic properties, graphic representation of data is preferred or at least needed, due to its intuitiveness and the possibility to provide unexpected discoveries that may be abstract and could not be quantified algorithmically.

However, unlike confirmatory data analysis which is built upon rigorous statistical procedures, e.g., hypothesis testing, visual diagnostics relies on graphical perception - human's ability to interpret and decode the information embedded in the graph (Cleveland and McGill, 1984), which is to some extent subjective. Further, visual discovery suffers from its

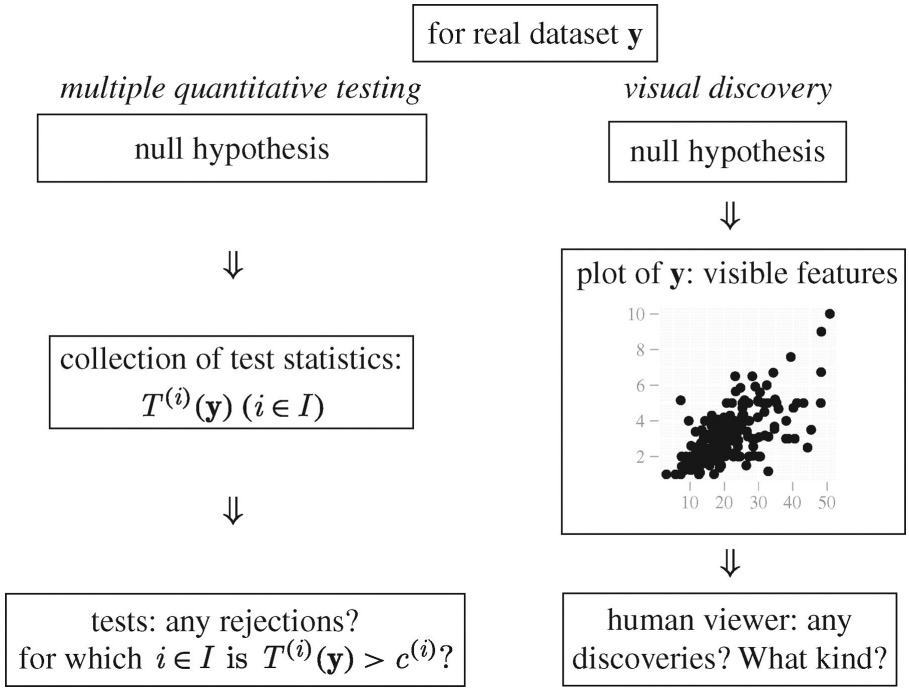

Figure 2.1: *Parallelism between multiple quantitative testing and visual discovery (Buja et al., 2009).*

unsecured and unconfirmed nature where the degree of the presence of the visual features typically can not be measured quantitatively and objectively, which may lead to over or under-interpretations of the data. One such example is finding a separation between gene groups in a two-dimensional projection from a linear discriminant analysis where there is no difference in the expression levels between the gene groups (Roy Chowdhury et al., 2015).

2.2.2 Visual Inference

Visual inference was first introduced by Buja et al. (2009) as an inferential framework to extend confirmatory statistics to visual discoveries. This framework redefines the test statistics, tests, null distribution, significance levels and p-value for visual discovery modelled on the confirmatory statistical testing. Figure 2.1 outlines the parallelism between conventional tests and visual discovery.

In visual inference, a visual discovery is defined as a rejection of a null hypothesis, and the same null hypothesis can be rejected by many different visual discoveries (Buja et al., 2009). For model diagnostics, the null hypothesis would be the assumed model, while the

visual discoveries would be any findings that are inconsistent with the hypothesis. The same assumed model, such as classical linear regression model, can be rejected by both nonlinearity and heteroskedasticity with the residual plot as shown in Figure 2.2.

2.2.3 Pre-specification of Visual Discoverable Features

As discussed in Buja et al. (2009), in the practice of model diagnostics, the range of possible visual discoveries is not pre-specified. In other words, people do not explicitly specify which one or more visual features they are looking for before the read of the diagnostic plot. This is concerning since conventional hypothesis testing always requires the pre-specification of the parameter space Θ of the parameter of interest $\theta \in \Theta$ to form a valid inferential procedure. To address this issue, a collection of test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ is defined, where $\mathbf{y}$ is the data and $\boldsymbol{I}$ is a set of all possible visual features. Buja et al. (2009) described each of the test statistics $T^{(i)}(\mathbf{y})$ as a measurement of the degree of presence of a visual feature. Alternatively, Majumder, Hofmann, and Cook (2013) avoids the use of visual features and defined the visual statistics $T(.)$ as a mapping from a dataset to a data plot. Both definitions of visual test statistics are valid, but in the rest of the paper, the first definition will be used as it covered some details needed by this work.

The size of the collection $T^{(i)}(\mathbf{y})$ $(i \in I)$ depends on the size of the set I. If one can define I comprehensively, i.e, pre-specify all the visual discoverable features, the validity issue will be solved. Unfortunately, to our knowledge, there is no such a way to list all visual features. In linear regression diagnostics, possible visual features of a residual plot may be outliers, shapes and clusters. But these are very vague features which does not cover the full list of visual features.

Similarly, Wilkinson, Anand, and Grossman (2005) proposed the work called graph theoretic scagnostics, which adopted the idea of "scagnostics" - scatter plot diagnostics from . It includes 9 computable scagnostics measures defined on planar proximity graphs: "Outlying", "Convex", "Skinny", "Stringy", "Straight", "Monotonic", "Skewed", "Clumpy" and "Striated" which attempts to describe outliers, shape, density, trend and coherence of the data. This approach is inspiring but it does not contain the complete list of visual

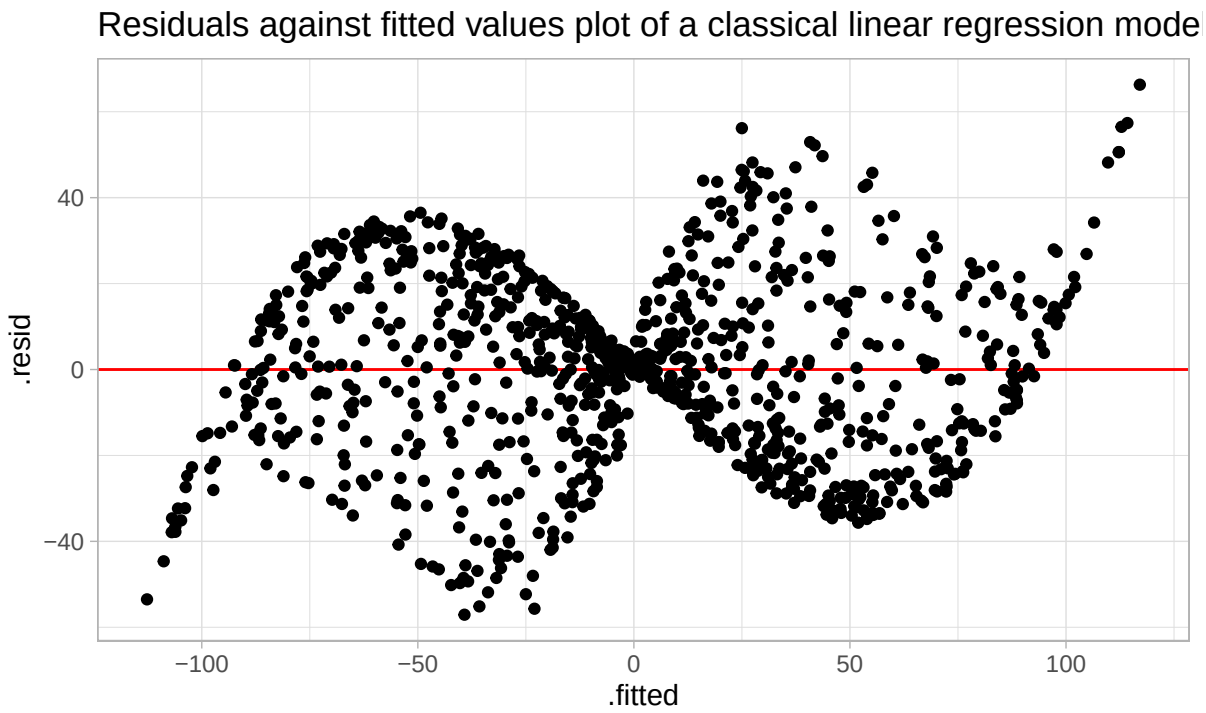

Figure 2.2: *test*

discoverable features. It is possible that such a list will never be complete as suggested in Buja et al. (2009).

Thinking out of the box, Buja et al. (2009) argued that there is actually no need for pre-specification of visual discoverable features. In model diagnostics, when the null hypothesis is rejected, the reasons for rejecting the hypothesis will also be known. This is because observers not only can point out the fact that visual discoveries have been found, but also can describe the particular features they observed. Those features will correspond to the subset of the collection of visual test statistics $T^{(i)}(\mathbf{y})$ $(i \in I)$ which resulted in rejection.

2.2.4 Lineup Protocol

Type-I error

Type-II error

2.2.5 Visual Inference Applied to Linear Regression

How people used visual inference in linear regression?

2.2.6 Limitation of the Visual Inference

What are the limitations?

2.2.7 Computer Vision Model

What is computer vision model?

2.2.8 Contribution

What has been done by this paper?

2.2.9 Structure of This Paper

What is the structure of the paper?

Model diagnostics is the part of data analysis whose primary objectives are to examine the goodness of the model fit and reveal potential violations of the assumptions. Graphical approaches

For regression diagnostics, it may includes the needs of

Linear regression is an modelling approach to describe the relationship between an response variable and one or more explanatory variable. It has been widely used for both generative modeling and predictive modelling.

Regression diagnostics is needed

1. to check whether the assumptions has been violated

2. to check whether the line fit the data

Model diagnostics for linear regression is well developed

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

Buja, A, D Cook, H Hofmann, M Lawrence, EK Lee, DF Swayne, and H Wickham (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906). Publisher: Royal Society, 4361–4383. (Visited on 01/24/2022).

Cleveland, WS and R McGill (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association* **79**(387). Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1984.10478080, 531–554. (Visited on 01/25/2022).

Majumder, M, H Hofmann, and D Cook (2013). Validation of Visual Statistical Inference, Applied to Linear Models. *Journal of the American Statistical Association* **108**(503). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2013.808157, 942–956. (Visited on 01/21/2022).

Roy Chowdhury, N, D Cook, H Hofmann, M Majumder, EK Lee, and AL Toth (2015). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. en. *Computational Statistics* **30**(2), 293–316. (Visited on 01/23/2022).

Wilkinson, L, A Anand, and R Grossman (2005). Graph-theoretic scagnostics. In: *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* ISSN: 1522-404X, pp.157–164.