Amendments Summary

Weihao Li

Changes made in response to comments from Examiner Prof. Louise Ryan

1. *While the candidate writes quite clearly, I feel that some improvements in presentation could be helpful. For example, in the abstract, it should be written in such a way that a non-expert could quickly get a sense of what the thesis is about and why it is important. In this case, I feel that the abstract is written in a fairly dry manner and it uses technical terms that are not appropriate without some context and definition. For example, I am a very experienced statistician, but I have never heard of the lineup protocol. Either this term should not be used in the abstract or there needs to be some kind of contextual statement describing what it is. The content of the thesis is actually very exciting. However, the abstract does not manage to communicate that excitement.*

I have revised the abstract to ensure that it provides a clear and concise overview of the thesis in a manner accessible to a broader audience, including non-experts. Technical terms have been either omitted or explained briefly with contextual statements to ensure clarity for readers unfamiliar with these concepts. For example, I have included a sentence explaining that the lineup protocol is a statistical visualisation approach that embeds a data plot among null plots to compare patterns arising by chance with those observed in the data.

2. *In Chapter 1, I can't help but feel that there should be more cited literature. The field of regression diagnostics is massive so I don't feel it is quite appropriate to sum it all up with a couple of references from the 80s and 90s! It may be adequate to say something along the lines that there is a vast literature out there and that more will be discussed in Chapter 2.*

To address the concern about summarizing the vast body of work, I added a statement in Chapter 1 acknowledging the extensive literature in regression diagnostics. I also clarified that this is a well-established field and noted that a more detailed discussion of the literature will follow in Chapter 2.

3. *related to the above point, it is probably wise to mention that regression diagnostics have been explored in a very wide range of contexts, including ordinary linear regression, generalised linear models, survival models, repeated measures models etc etc. I think*

it is correct to say that your thesis will be focussed on diagnostics for ordinary linear regression, is this correct? I think it is important to establish that context.

I have added a statement in Chapter 1 to highlight the wide-ranging applications of regression diagnostics across various models. I also clarified that the primary focus of my thesis is on diagnostics for classical normal linear regression model to establish this context early on.

4. *when describing the lineup protocol at the beginning of Chapter 1, perhaps some diagrams to help visualise how it works? As mentioned above, I have not myself heard of the lineup protocol, but it sounds very interesting and intriguing. Some visualisations might really help and also engage the reader. I know you do this in Chapter 2, but I can't help but feel it would be nice to see some kind of visualisation in the intro section as well.*

I have added a diagram in Chapter 1 to illustrate how the lineup protocol works, providing an example similar to those discussed in Chapter 2. This addition helps clarify its concept and engage the reader with the protocol from the outset.

5. *In chapter 2, can you say a bit more about how you recruited your subjects? I am not familiar with the Prolific Crowd Sourcing Platform. Were the subjects technically trained at all in terms of knowing what to look for?*

The recruitment details are provided in the appendix. The subjects received training by reading a webpage with examples. I decided not to modify this chapter further, as it has been published.

6. *I was glad to see a reference to Tukey in Chapter 3! He was absolutely one of the pioneers in terms of using visualisations to help interpret model fit etc. You might want to add some additional references earlier on as well. Some of his work was foundational in the field, in my opinion.*

I completely agree that Tukey's contributions were foundational and greatly impactful. While I appreciate the suggestion to add more references to his work earlier in the thesis, this one is sufficient to point to that body of work, and accommodate references to the range of foundational research on regression analysis.

7. *There are a few places where technical terms are used but not defined. e.g page 30, what is "dynamic input size"?*

I've addressed the issue by generally improving the language for better clarity in Chapter 3, including changing "dynamic input size" to "variable input size" and providing a clear definition on page 30.

8. *Perhaps it is there and I have not read carefully enough to understand, but in Chaper 2, are you assuming that the actual plot and all the null plots for the lineup protocol have been generated prior to being input into the computer vision model? If this is the case,*

then I agree it seems enough to enter residuals and fitted values. The only thing I would say is that sometimes there is interest in other sorts of plots, eg plots of residuals versus the values of some specific predictor variables.

Yes, the actual plot and all the null plots for the lineup protocol are generated prior to being input into the computer vision model. The chapter primarily focuses on residuals vs. fitted values plots. Other types of plots, such as residuals vs. specific predictor variables, are discussed in Section 3.12, which covers limitations and future work. No changes were made.

9. *related to the previous point, I found the term "model specification" (section 3.2) a bit confusing because I was thinking at first of the underlying model that you are trying to critique. Just be careful here to be clear since "model" is being used in a number of different ways.*

To avoid confusion, I have clarified the terminology by changing "model specification" in Section 3.2 to "computer vision model specification" This should make it clearer that the term refers specifically to the details of the computer vision model and not the underlying statistical model being critiqued.

10. *page 40, I don't see here where you talk about how you generate the values of x1 and x2?*

I have added a paragraph in Section 3.7 to provide more details on how the predictors x_1 and x_2 in the training datasets are generated.

11. *Figure 3.1 and the point made about it seems quite important. But the figure shows up a long way from where it is cited in the text*

The placement of Figure 3.1 was indeed a formatting issue in the original document. This has since been fixed, and the figure now appears much closer to its citation in the text.

Changes made in response to comments from Examiner Prof. Daniela Witten

1. *In Chapter 3, why not train a deep learning model to predict "contains outliers" versus "does not contain outliers", or "is non-linear" versus "is not non-linear", etc., rather than estimating the distance between the residual distribution and a theoretically "good" residual distribution? Wouldn't this obviate the need for the lineup evaluation in Section 3.5?*

This is something I considered carefully and discussed extensively with my supervisors during the development of the methodology. While training a deep learning model to predict categories like "contains outliers" vs. "does not contain outliers" or "is non-linear" vs. "not non-linear" seems appealing, it would bring us back to the traditional testing framework. These tests typically focus on detecting specific types of model violations in isolation.

One of the key advantages of using the lineup protocol for residual diagnostics is its ability to detect multiple model violations simultaneously, provided those violations produce noticeable differences in the data plots.

The lineup evaluation described in Section 3.5 remains valuable even in this context. By treating the model output as a statistic and comparing it to outputs generated under null assumptions, it adheres to the principles of visual inference and conventional statistical testing. This approach provides a robust framework for interpreting results and avoids the pitfalls of relying solely on predicted probabilities.

To clarify this, I have added two additional paragraphs in Section 3.2.2.

2. *In Section 3.5.1, it is mentioned that at least 100 null plots are needed to get a stable estimate of the 95th quantile. How many null plots are needed for this in Chapter 2?*

In Chapter 2, the data plot is compared against 19 null plots, as humans are not well-suited to evaluating lineups with a large number of plots. For reliable results, such a lineup should ideally be evaluated by at least five participants; otherwise, the smallest achievable p-value may not be sufficiently small. Additionally, involving multiple participants helps mitigate variability in individuals' abilities to assess residual plots. This approach is discussed in the experimental design and is supported by the cited literature in the chapter, such as Validation of Visual Statistical Inference, Applied to Linear Models by Mahbubul Majumder, Heike Hofmann, and Dianne Cook.

I decided not to modify this chapter further, as it has been published.

3. *In Section 3, the decision to treat this as a computer vision problem (rather than just passing in the vectors required to make a residual plot) was justified by saying that if the sample size is very very large, then it might not be possible to pass in the vectors that form the residual plot. I am a bit skeptical of this argument, since the case of an extremely large sample size, I think that the residuals could be sampled without replacement, with no loss of accuracy or ability to detect model violations. (Furthermore, I believe that ordinary least squares regression settings with such extraordinarily large sample sizes are relatively unusual — in contemporary settings, if the sample size is huge, then typically more complex models than ordinary least squares are fit.)*

The primary issue with providing pairs of vectors to the model is not the large sample size itself but the variability in sample size. Mainstream deep learning frameworks typically cannot handle inputs with variable dimensions. However, as you suggested, sampling pairs of fitted values and residuals without replacement could address this issue without sacrificing accuracy or the ability to detect model violations. This approach offers a way to define fixed-size inputs for the deep learning model.

I have updated Section 3.2.1 to include a discussion of this approach.

4. *How about a hybrid approach where someone visually screens a residual plot to identify possible violations (perhaps using the lineup method), and then conducts a formal statistical test only if possible violations are identified? The motivation here is that if a residual plot does not visually appear to have a violation, then it is not worth pursuing further. However, if a possible violation is visually identified, then a formal test may be worthwhile. This falls into the selective inference framework, since we are "double dipping" the data (once with a visual test, and then again with the formal test). Is it possible to conduct valid inference in this setting?*

This is indeed a valid inference setting and aligns with common practices in applications where findings from exploratory data analysis (EDA) are followed by formal statistical testing. Using visual methods, such as the lineup protocol, during EDA can help identify potential violations before formal testing.

However, this approach introduces the risk of multiple testing problems, as the data is effectively "double-dipped" - first for hypothesis generation and then for formal testing. Addressing this issue requires careful attention to selective inference frameworks. Possible solutions include splitting the data into distinct subsets for EDA and hypothesis testing or employing techniques like cross-validation to maintain robust inference.

That said, the primary challenge with this hybrid approach is its potential lack of practicality in many real-world scenarios. The goal of avoiding human evaluation is to minimize the workload for analysts. While the lineup method can be implemented with a simple line of code, it requires the additional effort of visually inspecting multiple plots, which can be time-consuming. In contrast, traditional tests are often preferred because they are easier to apply — typically requiring just a few lines of code or clicks in software interfaces.

As a result, this hybrid approach, while conceptually sound, may not be widely adopted due to its higher demands on time and effort. Given these considerations, I have not modified my thesis to incorporate this approach, as its practical feasibility is limited.