Automated Assessment of Residual Plots with Computer Vision Models

Response to reviewers

2026-02-20

We thank the editor and the reviewer for their constructive comments that have improved this paper. We have addressed all comments, colored in purple. In addition, we also include an annotated version of the paper (`diff.pdf`) with the difference between the original and revised versions produced using `latexdiff`.

Editor

One referee, an Associate Editor (AE), and I have reviewed your paper. Although the topic is interesting, the referee raised several major concerns, which are detailed in their reports. The AE also indicated in the report that "the description of the work and the work itself require additional detail and rigor." As such, I cannot accept your paper in its current form. Nonetheless, given its potential, I would consider a revision that adequately addresses all the referees' comments.

We thank the editor for their thoughtful assessment and for recognizing the potential of our work. We appreciate the constructive feedback provided by both the Associate Editor and the referee. In response, we have carefully revised the manuscript to address all major concerns, including clarifying the definition of the outcome variable for the computer vision models, explicitly describing the generation of training and test datasets, and detailing how model performance is assessed.

We have also improved the rigor and clarity of the descriptions throughout the paper, added additional tables and figures where needed, and revised sections to ensure that both the methodology and results are clearly presented. We hope that the revisions adequately address the issues raised and improve the overall readability and impact of the manuscript.

Reviewer

General Comments

The idea of the paper seems interesting: using computer vision to help identify residual plots which indicates model violation. For any computer vision task, the input is clearly the image. However, there should be a clear definition of the outcome variable that the algorithm wants to predict from images. From the current draft of this paper, it is unclear that what is the outcome variable that the computer vision is learning from the residual plots, how the training datasets are created for computer vision learning, and how the performance of the computer vision algorithm is assessed. Please see my detailed comments in items 10, 12,13. I suggest the authors make these basic setups explicitly defined from the very beginning of the article.

We thank the reviewer for the positive feedback and the constructive suggestions. We agree that a clear definition of the outcome variable, the training data generation process, and the performance assessment of the computer vision models is essential for clarity.

In response, we have revised the introduction to explicitly define these components from the outset. Specifically, we now clearly state that the computer vision model is a regression system that takes a residual plot as input and predicts a non-negative estimated distance. We also clarify that training and test datasets are generated from a synthetic data model simulating residual plots under various violation scenarios, with the target distances computed from the underlying data generating process, which quantifies the extent of model violations. Finally, we describe that model performance is evaluated on held-out test data and compared against classical residual diagnostic tests as well as results from a human-subject experiment.

Specific Comments

1. p3, l9: What is "the lineup protocol"?

The lineup protocol was described in third paragraph of the Introduction, but more information can be obtained in the cited reference in the paragraph. The lineup protocol is not the main focus of this paper, but it is important context. So we have added a brief description in the Introduction, including an example lineup (Figure 2) and provided a reference to the original paper for readers who need more background.

2. missing references: Loy and Hofmann (2013; 2014; 2015)

These three references are already cited in the manuscript. We are not sure if there was a misunderstanding, but we have confirmed that these references are present in the manuscript.

3. p8 l16: Why do we need to replace it by a full-rank covariance matrix?

We have added a new sentence to the corresponding paragraph, explaining that the covariance matrix is replaced with a full-rank diagonal matrix to ensure a non-degenerate distribution for KL divergence computation.

4. eq (2), typically KL is specified by KL(p—q) due to asymmetry

We have modified D_{KL} to $D_{KL}(P\|Q)$, thank you.

5. p9 l 53: to solve eq 2: using "evaluate" for "solve" is better

We have changed "solve" to "evaluate", thank you.

6. p19 l30: What is "the data generating process"? We don't know the true distribution of y.

The data generating process used for the human subject experiment is a special case of the process used to generate the training data for the computer vision model in this study. This process is described in Section 7, with additional details provided in the appendix A. We have also revised the text in Section 7 to clarify how the synthetic dataset is generated and why it is required. In addition, we have included Table 1 to summarise the training and test datasets.

In brief, for data observed in practice, the true distribution of y is unknown, and therefore the distance proposed in this study cannot be computed. In contrast, for synthetic data, we have precise control over the data generating process, so the true distribution of y is known. This allows us to generate pairs of residual plots together with their corresponding distances, which are used to train the computer vision model. This approach is a common practice for training deep learning models, as discussed in @nikolenko2021synthetic, which is now cited in Section 7.

7. p12, sec 5.1: this sounds like a standard simulation of the sampling distribution of $\hat{D}$ in traditional statistics.

We agree that our approach follows standard simulation-based methodology for approximating sampling distributions of a test statistic. To clarify this and address the reviewer's comment, we have revised the paragraph to explicitly acknowledge this connection (Paragraph 3, Section 5.1).

8. p13, Sec 5.2: How do you do bootstrapping? We need to know the distribution of $\hat{D}$ under the null. However, the given observed data y may not come from the null.

We use standard non-parametric bootstrapping, i.e., resampling the observed $\{(y_i, X_i)\}_{i=1}^n$ with replacement to form bootstrap samples, as stated in Section 5.2. For each bootstrap sample, the regression model is refitted and the corresponding residual plot is constructed, from which $\hat{D}$ is computed.

The null distribution of $\hat{D}$ is obtained separately from the bootstrap procedure. Specifically, assuming the fitted regression model is correct, we generate null residuals using residual rotation. These rotated residuals are used to construct null plots, which are then evaluated by the computer vision model to produce $\hat{D}_{null}$, forming an empirical approximation to the null distribution. In principle, this process could be repeated for each bootstrap sample to obtain a sample-specific $\hat{D}_{null}$. However, as discussed in Section 5.2, doing so would be computationally expensive.

We acknowledge that the observed y may not arise from the null model. In that case, the regression model is misspecified. If this misspecification is reflected in the residual plot, we expect the observed $\hat{D}$ (from the true residual plot) to be larger than many values from the null distribution, depending on the strength of the underlying signal. This is the central idea of the paper, and the corresponding testing procedure is described in Section 5.1.

Importantly, the bootstrap does not assume that the observed y comes from the null model. Its purpose is to quantify the sampling variability of $\hat{D}$ under the empirical data generating mechanism and to assess how frequently bootstrap replicates would lead to rejection. The null distribution used for inference is generated via residual rotation, not via bootstrapping.

9. Tbl1: it is unclear what this table is measuring. What is the R^2 measuring? What's the response and what is the predictors?

We have expanded the table caption to clarify what is being measured. The reported metrics evaluate the agreement between the estimated distance $\hat{D}$ (model output) and the target distance D generated from the synthetic data model. In particular, R^2 refers to the squared correlation between $\hat{D}$ and D.

10. Section 7 and 8: In computing $\hat{D}$, you need a P and Q for each targeted model violation. Is your computer vision learning algorithm targeted a particular model departure, eg, non-linearity or heteroskedasticity? What is your P and Q in generating the training data of $\hat{D}$ and residual plots for computer vision learning? However, your results also show your performance for different kinds of model violations. This is confusing. You need to specify clearly what is the "true" $\hat{D}$ and what the "predicted $\hat{D}$" using computer vision in generating training data of $\hat{D}$ and residual plots.

As stated in Section 7, "Importantly, the simulation included scenarios with multiple violations occurring simultaneously," the synthetic data generating process allows for both single and combined model violations. These violations include non-linearity, heteroskedasticity, and non-normality. The computer vision model is therefore not designed to target a specific violation in isolation, but rather to learn from the full spectrum of possible departures represented in the simulated data. To clarify this point, we have revised the text in Section 7 and added a table summarising the number of training and test samples across the different violation scenarios.

In computing $\hat{D}$, the reference distribution Q is always the residual distribution obtained by fitting a simple linear regression model under the standard assumptions, namely

$$Q = N(0_n, \text{diag}(R\sigma^2)).$$

The distribution P, by contrast, is determined by the synthetic data generating process for each scenario. It may reflect a single violation or a combination of violations, depending on the parameter configuration. The full specification of these synthetic models is provided in Appendix A.

Because the training data are generated under a range of violation settings, the resulting model performance can naturally be evaluated separately for different types of departures.

To further clarify the distinction between the D and the $\hat{D}$, we have added the following sentence to Section 7:

> *"The computer vision model was trained on paired data consisting of an RGB residual plot image as input and the corresponding target value D, with the objective of learning to estimate $\hat{D}$."*

Here, D denotes the distance computed from the known P and Q under the synthetic data generating process, while the computer vision model produces an estimate $\hat{D}$ from the residual plot image.

11. There is no numbering for your equations.

We have carefully reviewed all equations in the manuscript and confirm that the equation numbering follows the author guidelines: every equation that is explicitly referenced in the text is numbered, and equations that are not referenced are intentionally left unnumbered. During this check, we also verified that all in-text references correctly point to the corresponding numbered equations, so no further changes to the manuscript were required in response to this point.