RegLine: Assisting Novices in Refining Linear Regression Models

Xiaoyi Wang
University of Copenhagen
Copenhagen, Denmark
xiaoyi.wang@di.ku.dk

Luana Micallef†
University of Copenhagen
Copenhagen, Denmark

Kasper Hornbæk
University of Copenhagen
Copenhagen, Denmark
kash@di.ku.dk

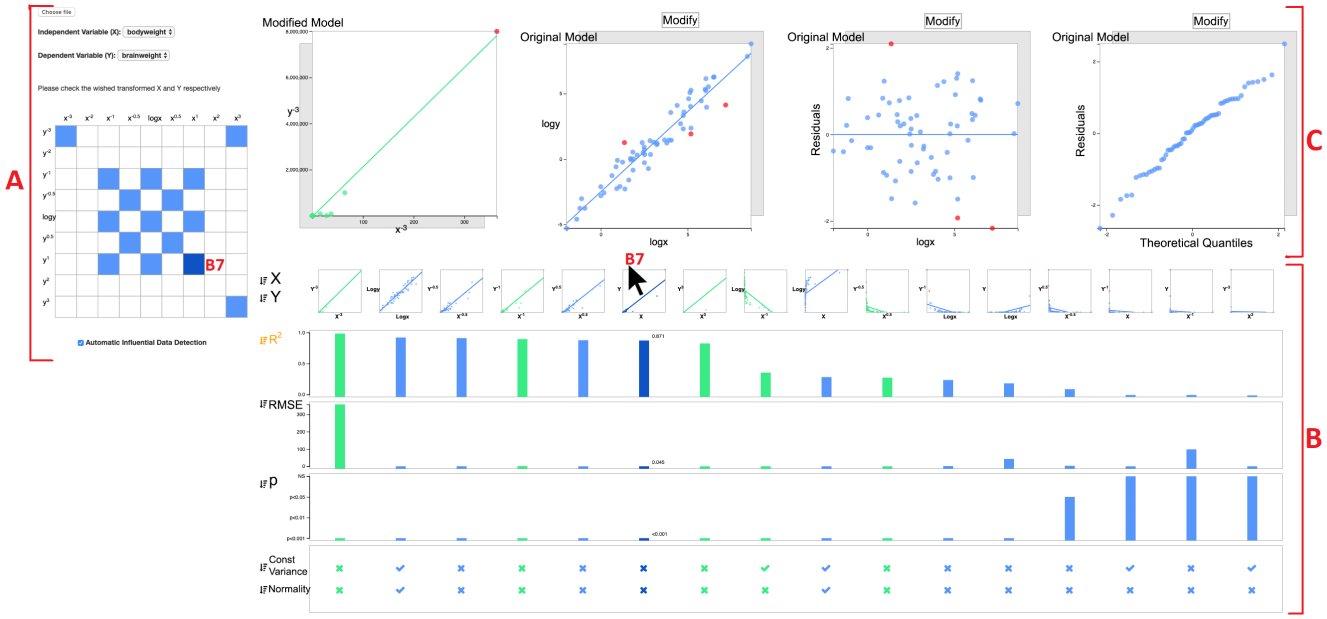

Figure 1: RegLine assists users in refining and validating simple linear regression models iteratively. Regline consists of three main components: (A) an input view for data transformation, (B) an overview of models for model comparison, and (C) a detailed view for residual analysis. B7 represents the highlighted model components (dark blue) by brushing and linking, when the cursor is over the model component. The light blue color represents the unchanged/original model, and the light green color represents the changed/modified model. The red dots represent unusual data points.

ABSTRACT

The process of verifying linear model assumptions and remedying associated violations is complex, even when dealing with simple linear regression. This process is not well supported by current tools and remains time-consuming, tedious, and error-prone. We present RegLine, a visual analytics tool supporting the iterative process of assumption verification and violation remedy for simple linear regression models. To identify the best possible model, RegLine helps novices perform data transformations, deal with extreme data points, analyze residuals, validate models by their assumptions, and compare and relate models visually. A qualitative user study indicates that these features of RegLine support the exploratory and refinement process of model building, even for those with little statistical expertise. These findings may guide visualization designs on how interactive visualizations can facilitate refining and validating more complex models.

† In Memoriam, February 2019.

AVI '20, September 28-October 2, 2020, Salerno, Italy

ACM ISBN 978-1-4503-7535-1/20/09…$15.00
https://doi.org/10.1145/3399715.3399913

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Visual analytics**; **Visualization systems and tools**.

KEYWORDS

model verification and remedy, linear regression, data transformation, residual analysis, exploratory data analysis

ACM Reference Format:
Xiaoyi Wang, Luana Micallef, and Kasper Hornbæk. 2020. RegLine: Assisting Novices in Refining Linear Regression Models. In *International Conference on Advanced Visual Interfaces (AVI '20), September 28-October 2, 2020, Salerno, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3399715.3399913

1 INTRODUCTION

Violating model assumptions and neglecting unusual data are the most common threats to the robustness of a linear model. To avoid such threats, users need to iteratively refine the model by techniques (e.g., data transformation, unusual data identification). However, the process of model refinement is not simple even for simple linear regression, because model refinement is a trial and error approach, which is time-consuming, tedious, and error-prone. It requires users to repeatedly go back and forth to tweak the model. It is difficult for novices to do this and even troublesome for experienced users. In particular, to fix the violations of model assumptions (e.g., linearity, normality, homogeneity), users may try a new transformation in each trial, and subsequently they need to (1) recheck all statistics, residual analysis, linear assumptions, and unusual data, (2) compare the current model with models from prior iterations. This cyclical process is continued until an appropriate and usable model has been found.

Current off-the-shelf tools including both programming-driven (e.g., R, MATLAB) and GUI-driven (e.g., SPSS) tools do not facilitate this iterative process of model refinement; they may make it even error-prone and inconvenient. For example, R users either change the same code or duplicate the code in each iteration to tweak the models. Furthermore, current tools still assume a certain level of statistical expertise from their users. For instance, R, MATLAB, and SPSS expect the user to know about and check specific model assumptions explicitly and provide no assistance on how the data could be transformed for the model to fit the data better. Additionally, current tools only allow the user to analyze one model at a time, so comparison or backtracking to previous models is hard or impossible.

Recent studies [8, 10, 14, 21, 23] mainly focus on the aspect of feature selection—steering multiple linear models with different subsets of independent variables and finding the optimal subset of variables among such subsets. Nevertheless, how interactive visualizations can facilitate verifying model assumptions and remedying the violations in the cyclical process still remain unexplored.

Therefore, we present RegLine, a visual analytics solution, to support this process of iterative refinement for simple linear regression models by exploratory data analysis. RegLine (Fig. 1) empowers non-statisticians in finding the right model for their data by: (1) exposing the effects of different transformations of data; (2) supporting the user in identifying potential influential data points (e.g., outliers, high leverage points) and in further investigating the effect of these points on the model's accuracy; (3) verifying that the data satisfies all of the model assumptions to ensure its robustness; (4) validating the accuracy of the models with respect to the necessary statistical tests, and showing how these accuracy measures relate; and (5) allowing visual comparison of different models through direct manipulation, brushing, and linking.

We first conduct an analysis of the tasks that users need to accomplish in fitting a simple linear regression model to their data sets and extract a set of requirements to support the exploratory model fitting and refinement process. We then devise RegLine and implement features from (1) to (5) to support the requirements. Finally, we evaluate RegLine for fitting simple linear regression models to real-world data sets in a qualitative user study.

Our evaluation demonstrates that RegLine's features support the process of model refinement and validation. We hope that this work will instigate further ideas on how interactive visualizations could empower non-statisticians (e.g., researchers from non-statistical domains) into refining and validating more complex models.

2 SIMPLE LINEAR REGRESSION ANALYSIS

Simple linear regression is a basic approach for supervised learning and it has been widely used for predicting a quantitative response by estimating the average value of y (the dependent variable or the predicted variable) given values of x (the independent variable or the predictor), as $y = \beta_0 + \beta_1 x + \epsilon$ [19]. Simple linear regression serves as a good jumping-off point for many fancy machine learning approaches which can be seen as generalizations or extensions of linear regression[16].

To estimate coefficients β_1 (slope) and β_0 (intercept) of the model, the least-squares estimator (LSE) is typically used [1], such that the random error ϵ (i.e., the residual sum of squares) is minimized. The model must then be evaluated as follows to determine the robustness and validity of the model in estimating the linear relationship between y and x (if any) [13]:

Linearity Assumption. The dependent variable y and the independent variable x should form a linear relationship–verified through visual analysis of the plot with the observed data.

Homogeneity Assumption. Residuals have equal variance–verified by Breusch-Pagan test [4].

Normality Assumption. Residuals are normally distributed–initially tested by Shapiro-Wilk test [26], but fully verified by a visual analysis of Q-Q (quantile-quantile) plot.

Independence. The residuals are independent–verified through visual analysis of the residual plot.

Unusual Data. Outliers and high leverage points can have significant influence on the regression analysis.

Model statistics. Coefficient of the determination $R^2 \rightarrow 1$, Statistical significance (ANOVA) of the slope $p < 0.05$, Standard deviation of prediction errors $RMSE \rightarrow 0$, such that the observed data points are close to the modeled line.

When a linear model violates one or more assumptions mentioned above, transforming x and y variables is the most frequently used method to remedy the violations while remaining within the simple or multiple linear regression framework. In addition to the logarithm transformation, there is a family of power transformations (e.g., Tukey's Ladder of Powers [30] or Box-Cox Transformation [3]) as a search space to explore.

Finding the right transformation for one or both of the variables y and x is often challenging. For instance, the class of transformations provided by Tukey's Ladder of Power [30] given by $z \rightarrow z^p$, whereby p is typically between -3 and 3, indicates 81 different possible transformations on both of the variables. Similarly, if influential

points are confirmed in the observed data, a new model excluding these points should be explored. If statistical summaries are not as robust as expected, the model is not an accurate estimate of the observed data and alternatives models need to be explored.

Every time a new model is explored, all of the above-mentioned model robustness and accuracy tests have to be verified. If more than one robust model is found during this exploratory data analysis, the best possible model should be determined through an in-depth comparative analysis of all of the model's accuracy statistics.

3 REQUIREMENT ANALYSIS

The design of RegLine is based on a set of requirements. To extract the list of requirements, we completed a review of the literature [1, 7, 9, 13, 15, 16, 19, 29] to understand data analysis tasks, challenges, and mistakes during the construction of a valid linear regression model. After a number of iterations of Munzner's nested model for visualization design and validation [22], we concluded the following list of requirements and categorized them according to model refinement (**R1, R2**), model validation (**R3, R4**), and model comparison (**R5, R6**):

R1: Show possible data transformations (Showing Transformations). Users should be provided with an overview of all possible transformations. Furthermore, each of the transformations would generate a different model. The user should be able to easily relate all of these models based on their similarity with respect to the type of transformations used.

R2: Unusual Data Identification. The user should be encouraged to further explore potential influential points in the observed data. Tentatively excluding data points can help users explore whether such points have great influence on the model.

R3: Visualize observed data and its modeled line (Visualizing Data and Line). Users can visually analyze unexpected aspects of the data and decide whether to explore alternative models with one or more of the variables transformed or with influential points in the observed data excluded.

R4: Visualize statistics of model robustness and accuracy (Visualizing Statistics and Residuals). Users should have access to detailed plots showing residuals and quantiles, together with the respective R^2, *RMSE*, and p. These statistics should be salient to the user, such that visual analysis of the model robustness and accuracy is encouraged.

R5: Rank Models. By allowing the user to rank the models by the different robustness and accuracy statistics, the user can better understand how different models relate to one another and which models should be investigating further.

R6: Allow a detailed comparison between models (Detailed Comparison). The user should be allowed to perform an in-depth comparative analysis of two models with respect to their estimated line and all of their robustness and accuracy statistics.

4 RELATED WORK

Many tools emphasized either visual analysis or statistical computing, but few of them combined the strength of both. For example, tools (e.g., Kinetica [25], TouchViz [11], EvoGraphDice [6], Tableau) demonstrated that direct manipulation can facilitate visual analysis, but they lack the power of statistical computation. In contrast, tools like Statwing [28], Wizardmac [31], R, MATLAB are strong at computational modeling, but direct manipulation and interactive visualizations are missing. Kehrer et al. [18] explored the use of interactive brushing to select subsets and dynamically exchange the subsets and summary statistics between R and visualization. RegLine steps further to investigate the potential of integrating statistical computing using R with the advantages of visual analytics.

Several studies have used visual analysis approaches to facilitate linear regression analysis [8, 10, 12, 14, 17, 21, 23, 32]. For example, Guo et al. [14] assisted users to discover linear trends among multiple variables by extracting subsets of independent variables. Similarly, Muehlbacher and Piringer [21] explored relationships between a feature space of independent variables and a target dependent variable by the partition-based framework. The framework helped users find an optimal subset of independent variables to build the regression model. RegressionExplorer [10] helped domain experts perform logistic regression by trying different subsets of independent variables. Besides, BEAMES [8] allowed domain experts to steer and inspect multiple different types of regression models. Although these studies mentioned above [8, 10, 14, 21, 23] do facilitate the iterative model refinement process, they mainly focus on the aspect of feature selection. That is, multiple models with different subsets of independent variables are iteratively steered to discover the most optimal subset of independent variables.

However, the question of how interactive visualizations can facilitate the aspect of model assumption verification and its related remedies seems to be unexplored in these studies. Our approach attempts to answer this question by integrating exploratory methods (e.g., data transformation, unusual data identification, residual analysis) into visualizations. A major difference between the existing studies and our approach is the levels of granularity in terms of the model refinement and validation. That is, the existing studies stop after one or more optimal models are found from multiple models with different subsets of independent variables. Nevertheless, our approach starts from their ending point by further checking the validity of each optimal candidate and remedying the violations. We focus on a model with a fixed subset of independent variables and turns this model to multiple models by tweaking data transformation and outliers.

Also, the fully automatic approach seems to be more promising and convenient. However, this approach may lead to the over-fitting problem [13]. For example, high transformation powers can make the statistics and the graph look appealing but bad performance in test data [16]. Algorithms can only select the best model based on statistics. Similarly, Cook's Distance method [7] only detects *potential* influential points. Thus human involved visual analysis is expected to determine whether such data points are really influential or not and whether the selected model is over-fitted or not. Furthermore, RegLine also gets novices engaged into the learning process of linear regression analysis.

Therefore, we start with simple linear regression as the first step to explore this field.

5 REGLINE

RegLine is a visual analytics solution that facilitates the refinement and validation of simple linear regression models. RegLine allows

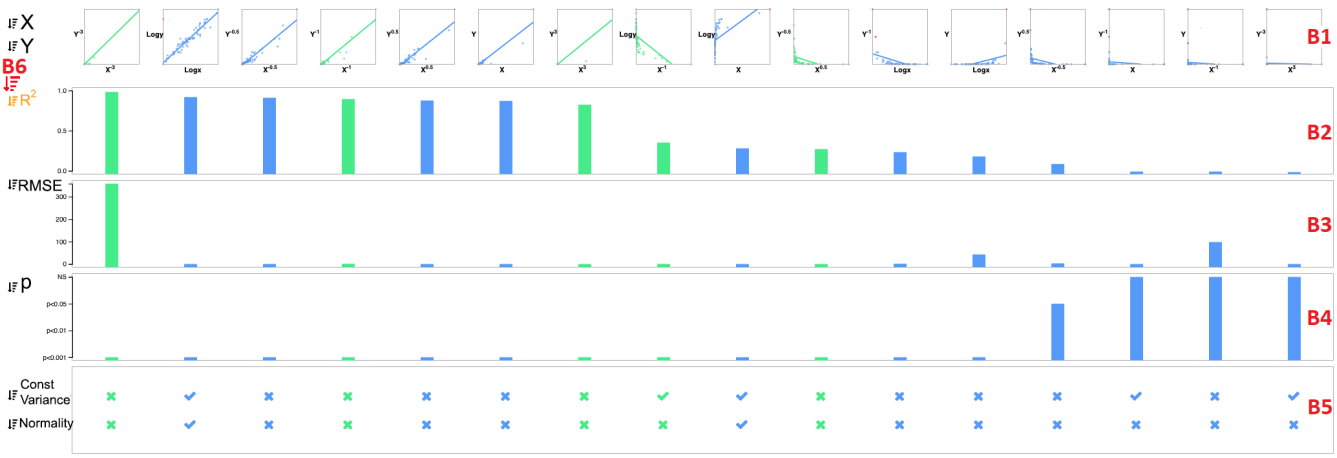

Figure 2: The overview consists of (B1) models in small multiples, (B2) bar charts of R^2, (B3) bar charts of RMSE, (B4) bar charts of p-value, (B5) plots of model assumptions, and (B6) model ranking. The light blue color represents the unchanged/original model, and the light green color represents the changed/modified model.

users to explore the linear models by both exploratory data analysis (EDA) and confirmatory data analysis (CDA). EDA helps users find patterns and gain preliminary evidence from data, and CDA assists users to draw conclusions based on the traditional statistical tools (e.g., significance, RMSE) [2]. The two analysis methods are neither mutually exclusive nor performed one after another but always complement each other to help users build the best possible model. RegLine consists of an input view (Fig. 1-A), an overview of models (Fig. 1-B), and a detailed view (Fig. 1-C), as shown in Fig. 1. We now discuss the design of these components and features of RegLine.

5.1 Input View

The input view (Fig. 1-A) allows users to formulate a linear relationship by specifying dependent y and independent x variables, make a set of transformations, and select the criteria by which to highlight influential points in the data.

5.1.1 Matrix of Data Transformation. RegLine addresses **R1-Showing Transformations**, which concerns possible ways of transforming the independent and dependent variables. Tukey's Ladder of Powers for data transformation is used in RegLine, which enables users to explore a linear relationship between x and y in a form of $y^{\lambda_1} = \beta_1 x^{\lambda_2} + \beta_0$ by adjusting the values of λ_1 and λ_2. This method is easy to understand for novices compared to Box-Cox Transformation, and it fits the scope of linear regression analysis.

A matrix (Fig. 1-A) is designed as a search space for users to explore possible combinations of x and y transformations. Some systems [10, 32] also employed the matrix approach and mapped two types of variables to the rows and columns of the matrix. For example, Dingen et al. [10] use rows for different models and columns for covariates. In contrast, RegLine maps three variates to the rows, columns, and cell grids of the matrix. Each column represents possible x transformations, each row indicates possible y transformations, and each grid cell represents a combination of x and y transformations. RegLine treats each combination of x and y transformations as a model. The matrix lists often used transformations of x and y (the exponent ranges from -3 to 3) [13, 19, 30]. Users are able to select or deselect a model by clicking. Colors are used to distinguish whether a model is selected (light blue) or not (white). The most frequently used transformations (-0.5, 0, 0.5, 1) [13] are given as default options, marked in blue. Previous studies show the possibility of multi-model steering technique which enables the speedy tweaking parameters with minimum input from the user [8, 20]. With the design of the matrix, RegLine allows users to steer multiple models synchronously by tweaking the parameter λ with multiple values. Users can dynamically add or delete models by the transformation matrix during the model refinement process.

Besides, instead of observing the original $x - y$ plot first or giving hints of how to transform, the matrix design lets users directly explore the space of all possible transformations and find trends/patterns. It also allows novices to learn why and how the transformations need to be performed through RegLine.

5.1.2 Highlight Potentially Influential Data Points. The LSE technique used in the simple linear regression modeling is very sensitive to extreme values of both x (high leverage points) and y (outliers) [1]. The lack of diagnostic influential observation analysis can severely affect the performance of LSE. RegLine detects potential influential data observations by Cook's Distance method [7], where an observation with Cook's distance larger than three times the mean Cook's distance might be influential to the regression line [13]. Highlighted potential influential points are marked as red (e.g., Fig. 3). Users can switch off this feature if they wish to manually discover influential points (Fig. 1-A).

5.2 Overview of Models

The overview of models (Fig. 2) consists of small multiples and statistics plots. Users can observe an overview of all models, model statistics, and compare models across different measures.

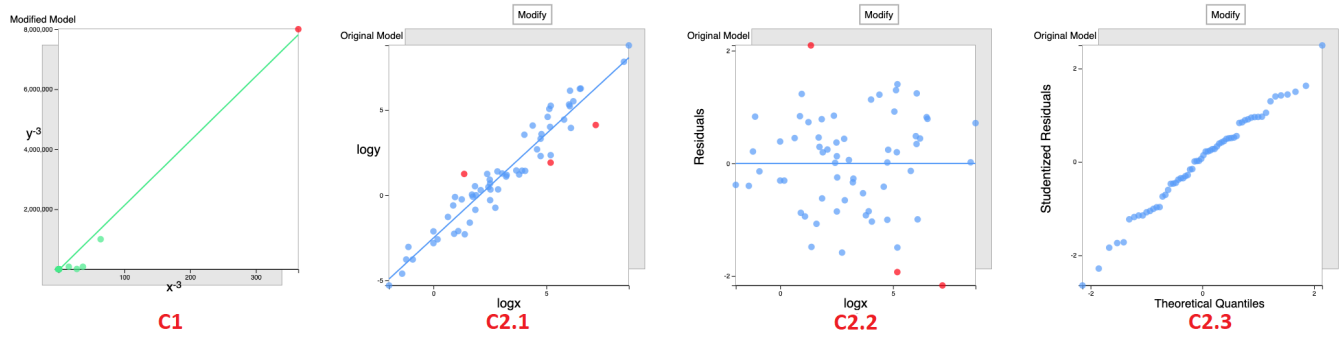

Figure 3: The detail view consists of (C1) reference view, (C2.1) data view, (C2.2) residual plot, and (C2.3) Q-Q (quantile-quantile) plot. The light blue color represents the unchanged/original model, and the light green color represents the changed/modified model.

5.2.1 Small Multiples. We use small multiples (Fig. 2-B1) to show all selected models from the transformation matrix (Fig. 1-A) because this design allows users to have a quick grasp of the overview of different data transformation effects (**R1-Showing Transformations**). Each of the small-multiple charts corresponds to a selected model with a combination of the x and y transformations in the matrix (Fig. 1-A). Each small-multiple chart consists of a scatter plot representing data and a regression line produced by LSE (**R3-Visualizing Data and Line**); users can compare different transformations across models.

5.2.2 Statistics of Model Robustness and Accuracy. Bar charts are the most effective magnitude channel for ordered attributes [22] compared to other types of charts, so RegLine quantifies R^2, RMSE, and the p-value of the slope β_1 by bar charts to depict the statistics of the model accuracy. Fig. 2-B2, Fig. 2-B3, and Fig. 2-B4 represent the three model statistics R^2, RMSE, and p-value of the slope β_1 respectively. RegLine uses checkmark and cross symbols to represent whether models fulfill the normality and homogeneity assumptions in Fig. 2-B5. RegLine addresses **R4-Visualizing Statistics and Residuals**. Also, each small-multiple chart and its corresponding bars are placed in the same column. Users can link them to the corresponding grid cell in the matrix by brushing and linking (e.g., Fig. 1-B7 highlighted by dark blue).

5.2.3 Ranking Models. Users can rank all models decreasingly with respect to different statistics (i.g., R^2, RMSE, p-value, normality, constant variance, transformation exponent) by clicking the ranking icon (Fig. 2-B6). This ranking addresses **R5-Ranking Models**. R^2 is set as the default sorting option. The ranking feature allows users to relate models produced by transformation exponents visually (**R1-Showing Transformations**). Users can explore how transformation exponents influence models and reason about fitted transformations by trends reflected in small multiples.

5.3 Detailed View

The detailed view (Fig. 3) allows users to explore models in depth, make changes on models, compare two different models.

5.3.1 Exploratory View. The exploratory view assists users with influential data identification and residual analysis. The view consists of three plots—the data plot (Fig. 3-C2.1), the residual plot (Fig. 3-C2.2), and the Quantile-Quantile (Q-Q) plot (Fig. 3-C2.3). The view is triggered by clicking on a small-multiple chart. The three plots are the detailed views of the corresponding model and depict different aspects of the model. The data plot in Fig. 3-C2.1 is an enlarged small-multiple chart, and it depicts the correlation between transformed x and y. Users can find patterns from the data, validate the linearity assumption, and identify unusual observations by observing the data plot (**R3-Visualizing Data and Line**).

The residual plot in Fig. 3-C2.2 assists users with the residual analysis. Users can validate the homogeneity assumption and the independence assumption, look for patterns, and identify influential observations from the residual plot (**R4-Visualizing Statistics and Residuals**). The Q-Q plot in Fig. 3-C2.3 helps users reason about whether the residuals are normally distributed. The linearity of the points in the Q-Q plot indicates the normality of the residuals. Q-Q plot is visually more intuitive, and it also gives the exploration space when the statistical result is marginally significant.

5.3.2 Switch States. RegLine allows users to modify the model but keep the original model as a reference for the later comparison. Therefore, RegLine stores two states for each model. One state is the original state, and the other state is the modifiable state. Users can only modify a model in the modifiable state. To reflect this design on visualizations, each of the three plots consists of an original model view and a modifiable model view. Fig. 3-C2.1 shows that two views are overlapped and users can switch their interested view to the front by clicking the grey area. When a view is switched to the front, all its associated bars and charts are also switched along with the view. To better distinguishing two views, we use the color encoding to distinguish the unchanged and modified models, so we choose similar hues—light blue and light green to show the closeness of the two models. Fig. 4 shows the two colored models. The blue represents the original model, and the green represents the modifiable model.

5.3.3 Modification Mode. We design a mode for model modification, which only allows users to modify one plot at a time. To enter this mode, users need to click the "modify" button above the

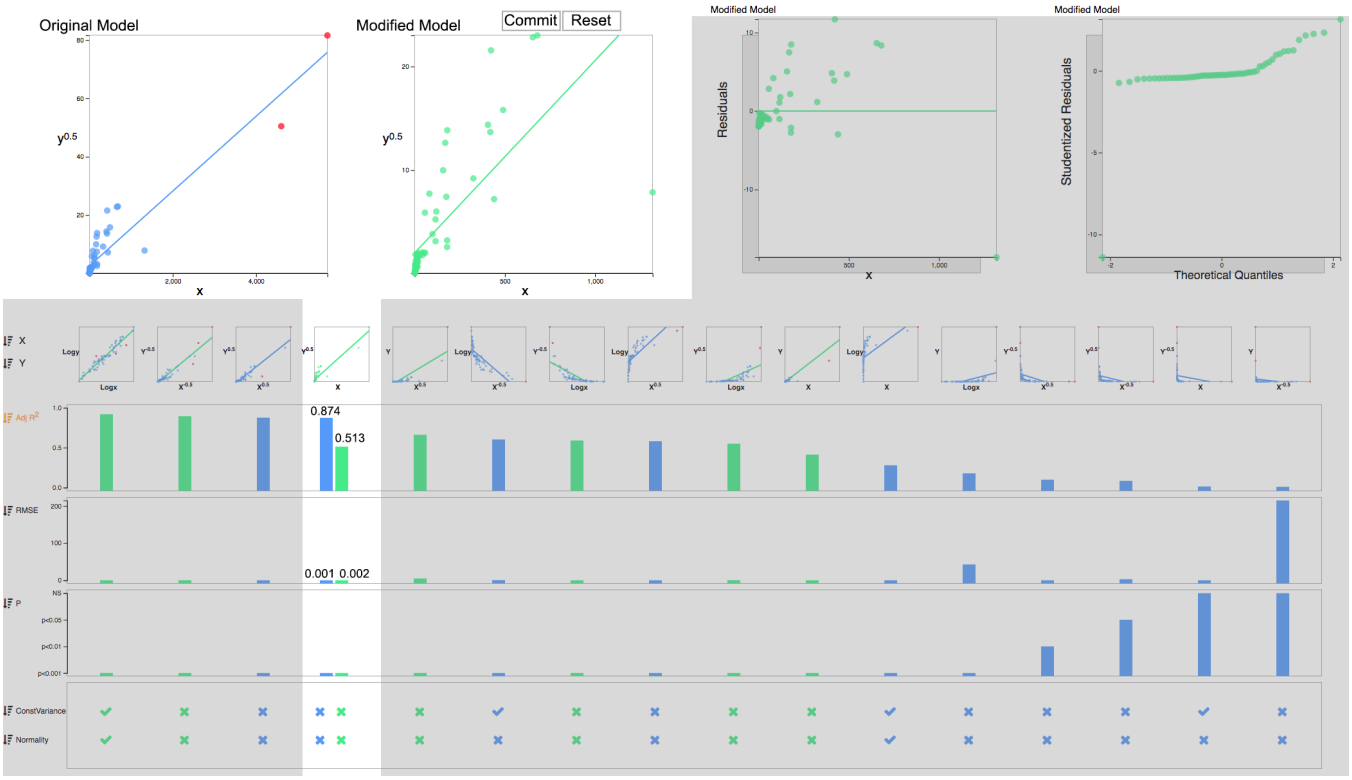

Figure 4: Modification Mode. Users can only interact with enabled plots in this mode. The disabled plots are displayed in gray. The light blue color represents the unchanged/original model, and the light green color represents the changed/modified model.

wished plot. Fig. 4 shows that other parts of RegLine are disabled in this mode and are colored to light grey. Only the original view, the modifiable view and the corresponding bar charts are enabled. Users can remove data points, observe the influence visually in the modifiable view, and compare the modified model with the unchanged model. In addition to the comparison between two views, RegLine also facilitates comparison in the model statistics plots in the modification mode. Users can observe the immediate update to the statistics of the modifiable models. This mode feature addresses **R6-Detailed Comparison**.

5.3.4 Potential Influential Data Identification. Influential data points can severely influence the performance of the LSE. Therefore, analyzing influential observations is essential. RegLine allows users to identify potential influential observations and attempt to remove them by clicking in the modification mode (**R2-Unusual Data Identification**). However, the design of excluding data points does not aim to facilitate p-hacking but rather draw attention to the unusual data through the animated transition of the regression line. A great influenced regression line can lead to a further investigation of the influential data point. Similarly, a less affected regression line helps users label the data point as usual even if it is visually or algorithmically unusual. Meanwhile, the animated transition also serves for an educational purpose which makes novices better understand the influence of potential influential points and the importance of unusual data analysis.

5.3.5 Reference View. RegLine supports comparing two different models in detailed views (**R6-Detailed Comparison**). Fig.3-C1 shows a reference view which is set as a comparison with the exploratory view. Users can drag and drop a model from small multiples to the reference view. The reference view allows users to compare two different models side by side. However, the reference view only supports one type of plot (i.e., data scatter plot) and users cannot modify models in the reference view.

5.4 Comparison of Models

Model comparison is important for model selection. Several comparison techniques have been discussed in previous sections. In this section, we summarize comparison techniques from two angles.

Comparison between different models. RegLine allows users to compare the difference through small multiples to get an overview of models, through statistics plots and sorting, and through side-by-side technique (reference view vs. detailed view).

Comparison within the same model between original (all data) and modified (partial data) versions. In the modification mode, users can compare a modified model with its unchanged

model in a detailed view. Meanwhile, they can also compare statistics between the two models.

6 EVALUATION

6.1 Rationales

As a first step to explore supporting iterative model refinement and validation for linear models, we focused on gaining formative insights into how participants refine and validate simple linear regression models by using RegLine's features. Therefore, we opted for a think-aloud study to observe and understand the reasoning process of our participants, as they were using RegLine. A pilot study with two participants was conducted to verify the appropriateness of our study procedure and tasks.

6.2 Participants

Eight participants (6 males:2 females; age: [25-36], m=26.8, SD=3.7) were recruited. Participants were either university researchers or students with a background in computer science, bioinformatics, or geoscience. The average level of self-reported expertise on linear regression analysis noted by the participants on a 7-point Likert scale was 4 (1=no knowledge, 7=expert).

6.3 Task and Data

Participants were asked to build and refine simple linear regression models based on three real-world data sets, D1[5], D2[13], D3[27]. D1 was only for the training purpose. The order of D2 and D3 was counterbalanced between participants. The characteristics of D1-D3 captured the typical challenges in linear regression analysis: containing influential data points and violating the model assumptions of constant variance, normality, and linearity.

6.4 Procedure

We gave participants an introduction about how to use RegLine and asked the participants to use it and play with the data set D1 for 15 minutes. Next followed two trials. In each trial, a data set with a scenario was given to participants. After reading the scenario text, they were asked to think aloud while they were refining the models. We observed and took notes on their approaches to refining models and encountered problems. After participants completed the trials, we conducted a semi-structured interview with the participants. The average time of the whole study was 60 minutes. The introduction and training took 25 minutes on average. The total time of two trials was about 20 minutes. The interview was about 15 minutes per participant. RegLine was displayed on an external 27-inch monitor (Apple) with a 2560 x 1440 pixels resolution. Each participant was rewarded a gift to acknowledge their participation.

7 RESULTS

This study aimed to gain insights into how participants refine and validate simple linear regression models by using the features of RegLine. Based on the think-aloud study and interviews, we believe that there is initial evidence that RegLine's features support the model refinement and validation, even for novices. We summarized the findings as follows which may contribute to the future design of refining and validating more complex models.

Data Transformation. Participants used different approaches to exploring the possible combinations of x, y transformations and steering multiple models. Four participants did not use default transformation options by RegLine but attempted to identify the optimal model in their ways. For instance, P2-P4 first deselected all default options and started with extreme transformations on the corners of the matrix. Then, they gradually moved the search towards the center of the matrix. P6 selected all transformations on the matrix at the beginning and started exploring the high-ranking models. P1, P5, P7, and P8 began with the default transformations and gradually expanded the search area in the matrix. Although it is a trial-and-error approach, they attempted to reason about the best possible transformation based on the selected ones instead of randomly clicking.

Furthermore, participants felt that they had the flexibility to try different transformations (**R1-Showing Transformations**), which brought different insights to the data (e.g., "I like the freedom given by the transformation matrix and I can see the data from different angles... It gives a search space to explore different transformations...The tool is a combination of brute-force method and heuristic exploration to search for the best models", P2). Five participants mentioned that the transformation matrix showed the relationships between the transformations and models and allowed them to steer multiple models. Four participants further stated that this matrix not only gave them cues of which transformations they should explore next, but also prevented them for ending up in a local minimum. In addition, participants mentioned that RegLine could save them efforts and time compared to tools like R, Python (e.g., P3 "I can focus on the analysis without worrying about scripting").

Influential Data Identification and Auto-detection. All participants checked the check-box (Fig. 1-A) to highlight the influential points detected by RegLine. They attempted to remove highlighted points and observed the influence on the regression line through animations. Meanwhile, they also compared the original model and the modified model by the two regression lines and the bar charts for model statistics. In addition to the highlighted points, participants also attempted to remove the points that they felt suspicious. However, all participants were cautious with excluding data points. P2 and P4-P8 reset the modified model to the original one because they believed that the excluded points had little influence on the regression line. They treated them as usual points even if they were marked as unusual by the algorithm.

Meanwhile, they were also surprised that algorithms were not reliable as what they expected because of the false-positive unusual data. For instance, P1 stated "I have never questioned the algorithm, and the automated algorithm should always be right in my mind." Participants found the modification mode very useful which allowed them to explore potential outliers without ruining the data (**R2-Unusual Data Identification, R6-Detailed Comparison**). For example, P7 stated "with the reset function, I do not have to worry that data is messed up after the outlier removal." Furthermore, P2, P4, and P7 mentioned that the process of the tentative removal and reset made them learn the importance of analyzing unusual points instead of just excluding them.

Residual analysis and Assumption checking. During the interview, six participants mentioned that they do not perform residual analysis in their daily practice. They were not aware of the

importance of checking residuals and model assumptions. Three of them were not familiar with how to visualize residuals or how to run statistical tests for model assumptions.

However, all participants used the detailed view to perform residual analysis and check model assumptions in RegLine. They checked unusual data and the linearity assumption through the data plot. P2, P3, and P5 used the residual plot to validate both independence and constant variance assumptions. The rest of them only validated the constant variance assumption by the residual plot. They also checked the normality assumption through the Q-Q plot. Meanwhile, they compared their reasoning based on the plots with calculated model-assumption statistics in the bar charts.

Participants gave similar reasons why they checked those measures in RegLine (**R3-Visualizing Data and Line, R4-Visualizing Statistics and Residuals**). For example, one of the participants explained "I feel I need to check the assumptions and do residual analysis because the visualization draws attention to them. It seems that they are important measures of the model in the visualization. I should not ignore them, especially when I saw the red cross icons in the charts."

Model Comparison. Participants found the comparison between original and modified models useful (e.g., P6: "I can see how the regression line was influenced between two models.") (**R6-Detailed Comparison**). They treated small multiples as an overview of models and often made a quick comparison across the models before seeing their detailed views. They also saw small multiples as a connection between bar charts, detailed views, and the matrix. Besides, the participants compared model statistics by observing the bar charts. All of them kept the model assumptions in mind when comparing models.

Model Ranking. Participants used the ranking to quickly narrow down the scope of possible models by removing "wrong" models from RegLine. They sorted the models according to different features, but most of them took R^2 as the primary criterion. P2 and P3 even further reasoned about possible transformations according to ranked small multiples (**R5-Rank Models**). Furthermore, they validated models not only by statistics but also by visualizations (**R4-Visualizing Statistics and Residuals**). The two behaviors were intertwined. In general, participants found the model ranking feature useful. For instance, "It is quite useful when people are only interested in certain features" (P1) and "I think the ranking is useful and practical because it can sort data in different ways" (P4).

8 DISCUSSION AND FUTURE WORK

RegLine attempted to support the activity of iteratively refining and validating models through simple linear regression analysis. Our user study suggests that RegLine supports iterative model refinement and validation, but deserves to be further refined. Next, we discuss four areas of improvements.

Gaps between RegLine and different user groups. Our target user group for this type of support was users with little statistical knowledge and skills. However, expert users or statisticians may expect more advanced features and the flexibility to tweak algorithms and parameters of the models via RegLine. For instance, instead of ordinary least squares, weighted least squares can be an alternative for estimating the regression line. Also, the feature of ranking models can be extended to sort models by multiple features instead of a single feature. We should further explore how RegLine can be useful for a more broad group of users.

P-hacking. There is increasing concern about the p-hacking issue because users may manipulate data to produce desired p-values by visual analytics systems. However, we did not aim to facilitate p-hacking in RegLine but let novices understand the importance of influential observation analysis. Although all participants were cautious with excluding data, we still believe it is necessary to investigate systematical designs and guidance for avoiding the p-hacking issue. Furthermore, statistical tools (e.g., multiple comparison correction, regularization) and visual corrections could be integrated into the system in order to prevent this issue from propagating further down an analysis pipeline [24].

Scalability of RegLine. One limitation of RegLine is that at the moment it can only be used for simple linear regression analysis. Thus in the future, we need to investigate how RegLine could be extended to handle more general linear models (e.g., multiple linear regression). Moreover, RegLine only lists the most frequently used exponents of the x and y. Scaling to more data and more transformations need to be considered in further work. It would be essential to explore how visualizations could help users search and steer multiple models when the search space of transformations increases.

User Study. Our user study is a first step towards evaluating RegLine; focused on rich, qualitative data from a small set of users. Thereby, we cannot draw statistically valid conclusions about modeling performance. We cannot conclude strongly about the influence of statistical knowledge, either, because the small sample of users is relatively heterogenic in backgrounds. Thus, future work should evaluate RegLine with a large sample of users and do so in an experimental setup where users with knowledge in statistics may be compared to statistical novices.

9 CONCLUSION

Simple linear regression modeling is essential in many domains. Current statistical and visualization tools, however, do not support the refining and validating of such models well. This is particularly problematic in situations where novice users are exploring data to find an apt model. We have presented RegLine, a visual analytics solution for flexibly exploring, refining, and validating a multitude of models. RegLine satisfies a set of 6 requirements which we have extracted for linear regression model analysis. Our user study shows that RegLine helps the user in exploring and comparing models, their statistics and the various ways that the data corresponding to the model variables could be transformed. Other features in RegLine that were particularly useful to our participants indicate its ability to easily rank models based on their statistics, indicate and support the user in investigating possible influential data points, and highlight data sets that do not satisfy any of the model assumptions.

ACKNOWLEDGMENTS

The authors would like to acknowledge Innovation Fund Denmark and the BIOPRO2 strategic research center (Grant No. 4105-00020B).

REFERENCES

[1] Naomi Altman and Martin Krzywinski. 2015. Simple linear regression. *Nature Methods* 12 (Oct. 2015), 999. http://dx.doi.org/10.1038/nmeth.3627

[2] John Behrens. 1997. Principles and Procedures of Exploratory Data Analysis. *Psychological Methods* 2 (06 1997), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

[3] G. E. P. Box and D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26, 2 (1964), 211–252. http://www.jstor.org/stable/2984418

[4] T. S. Breusch and A. R. Pagan. 1979. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* 47, 5 (1979), 1287–1294. http://www.jstor.org/stable/1911963

[5] John Burkardt. [n.d.]. STATS - Statistical Datasets. http://people.sc.fsu.edu/~jburkardt/datasets/stats/stats.html http://people.sc.fsu.edu/~jburkardt/datasets/stats/stats.html, Mar. 2019.

[6] W. Cancino, N. Boukhelifa, and E. Lutton. 2012. EvoGraphDice: Interactive evolution for visual analytics. In *2012 IEEE Congress on Evolutionary Computation*. 1–8. https://doi.org/10.1109/CEC.2012.6256553

[7] R.D. Cook and S. Weisberg. 1982. *Residuals and Influence in Regression*. Chapman & Hall. https://books.google.dk/books?id=MVSqAAAAIAAJ

[8] Subhajit Das, Dylan Cashman, Alex Endert, and Remco Chang. 2018. BEAMES: Interactive Multi-Model Steering, Selection, and Inspection for Regression Tasks.

[9] David M. Lane. [n.d.]. Online Statistics Education: A Multimedia Course of Study. http://onlinestatbook.com http://onlinestatbook.com, Mar. 2019.

[10] D. Dingen, M. van't Veer, P. Houthuizen, E. H. J. Mestrom, E. H. H. M. Korsten, A. R. A. Bouwman, and J. van Wijk. 2019. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan 2019), 246–255. https://doi.org/10.1109/TVCG.2018.2865043

[11] Steven M. Drucker, Danyel Fisher, Ramik Sadana, Jessica Herron, and m.c. schraefel. 2013. TouchViz: A Case Study Comparing Two Interfaces for Data Analytics on Tablets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2301–2310. https://doi.org/10.1145/2470654.2481318

[12] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. 2011. Observation-level interaction with statistical models for visual analytics. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 121–130. https://doi.org/10.1109/VAST.2011.6102449

[13] John Fox. 2008. *Applied regression analysis and generalized linear models, 2nd ed.* Sage Publications, Inc, Thousand Oaks, CA, US.

[14] Z. Guo, M. O. Ward, and E. A. Rundensteiner. 2009. Model space visualization for multivariate linear trend discovery. In *2009 IEEE Symposium on Visual Analytics Science and Technology*. 75–82. https://doi.org/10.1109/VAST.2009.5333431

[15] Iain Pardoe, Laura Simon, and Derek Young. [n.d.]. STAT 501. https://newonlinecourses.science.psu.edu/stat501/node/2/ https://newonlinecourses.science.psu.edu/stat501/node/2/, Mar. 2019.

[16] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[17] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. 2009. iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum* 28, 3 (2009), 767–774. https://doi.org/10.1111/j.1467-8659.2009.01475.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2009.01475.x

[18] J. Kehrer, R. N. Boubela, P. Filzmoser, and H. Piringer. 2012. A generic model for the integration of interactive visualization and statistical computing using R. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 233–234. https://doi.org/10.1109/VAST.2012.6400537

[19] M.H. Kutner, C.J. Nachtsheim, and J. Neter. 2003. *Applied Linear Regression Models*. McGraw-Hill Higher Education. https://books.google.dk/books?id=0nAMAAAACAAJ

[20] Gonzalo Gabriel Méndez, Uta Hinrichs, and Miguel A. Nacenta. 2017. Bottom-up vs. Top-down: Trade-offs in Efficiency, Understanding, Freedom and Creativity with InfoVis Tools. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 841–852. https://doi.org/10.1145/3025453.3025942

[21] T. Muehlbacher and H. Piringer. 2013. A Partition-Based Framework for Building and Validating Regression Models. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec 2013), 1962–1971. https://doi.org/10.1109/TVCG.2013.125

[22] Tamara Munzner. 2014. *Visualization analysis and design*. AK Peters/CRC Press.

[23] H. Piringer, W. Berger, and J. Krasser. [n.d.]. HyperMoVal: Interactive Visual Validation of Regression Models for Real-Time Simulation. *Computer Graphics Forum* 29, 3 ([n. d.]), 983–992. https://doi.org/10.1111/j.1467-8659.2009.01684.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2009.01684.x

[24] X. Pu and M. Kay. 2018. The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*. 37–45. https://doi.org/10.1109/BELIV.2018.8634103

[25] Jeffrey M. Rzeszotarski and Aniket Kittur. 2014. Kinetica: Naturalistic Multi-touch Data Visualization. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 897–906. https://doi.org/10.1145/2556288.2557231

[26] S. S. Shapiro and M. B. Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 3/4 (1965), 591–611. http://www.jstor.org/stable/2333709

[27] Helmuth Spath. 1992. *Mathematical Algorithms for Linear Regression*. Academic Press Professional, Inc., San Diego, CA, USA.

[28] Statwing. [n.d.]. Statwing | Efficient and Delightful Statistical Analysis Software for Surveys, Business Intelligence Data, and More. https://www.statwing.com/ https://www.statwing.com/, Mar. 2019.

[29] Thomas C. Harrington. 2019. Statistical Methods for Management. http://ruby.fgcu.edu/courses/tharring/80890/

[30] John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley.

[31] Wizard. [n.d.]. Wizard: Statistics & Data Analysis Software for Mac. https://www.wizardmac.com/ https://www.wizardmac.com/, Mar. 2019.

[32] C. Zhang, J. Yang, F. Benjamin Zhan, X. Gong, J. D. Brender, P. H. Langlois, S. Barlowe, and Y. Zhao. 2016. A visual analytics approach to high-dimensional logistic regression modeling and its application to an environmental health study. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*. 136–143. https://doi.org/10.1109/PACIFICVIS.2016.7465261