Journal of Scientific Research and Reports

Volume 30, Issue 10, Page 250-266, 2024; Article no.JSRR.123452
ISSN: 2320-0227

Overview of Regression Models and How to Determine the Best Model for Data

N Madhavi Latha [a++], K. Geetha [b#] and S Damodharan [a*]

[a] Department of Statistics, Sri Venkateswara University, Tirupati, India.
[b] Department Mathematics Kanchipuram, Sri Sankara Arts and Science College, Tamil Nadu, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/jsrr/2024/v30i102452

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/123452

Review Article

Received: 14/07/2024
Accepted: 17/09/2024
Published: 25/09/2024

ABSTRACT

The selection of the appropriate regression model for your data is an essential stage that has a significant impact on the precision and interpretability of your analytical process. The purpose of regression models is to investigate the connections that exist between a dependent variable and one or more independent variables from a statistical perspective. The process of selection starts with gaining an awareness of the many kinds of regression models that are accessible. These models include linear, polynomial, and logistic regression, among others. Each of these models is suitable for a different kind of data and a different kind of connection. In order to reduce the number of possibilities, it is helpful to do an analysis of the features of your data, which may include the existence of outliers, multicollinearity, and distribution. Furthermore, each m0odel is accompanied by a set of unique assumptions, such as the linearity and normalcy characteristics that are

---

[++] Research Scholar;
[#] Assistant professor;
**Corresponding author: E-mail: damodharan.stat@gmail.com;*

Cite as: *Latha, N Madhavi, K. Geetha, and S Damodharan. 2024. "Overview of Regression Models and How to Determine the Best Model for Data". Journal of Scientific Research and Reports 30 (10):250-66. https://doi.org/10.9734/jsrr/2024/v30i102452.*

necessary for linear regression. In order to get findings that can be relied upon, it is essential to verify that these assumptions are correct; if they are not, different models such as generalized linear models would be required. In order to prevent overfitting, which occurs when the model captures noise rather than the actual data structure, it is essential to strike a balance between the complexities of the model. When it comes to making this judgment, methods such as cross-validation might be of assistance. In conclusion, it is important to take into consideration the trade-off between interpretability and predictive strength. Models that are simpler, such as linear regression, are simpler to explain, but models that are more complicated might produce better forecasts. You will be able to pick the regression model that is the most suitable for your data if you give careful consideration to these aspects, which will result in insights that are both robust and relevant. Selecting regression types depends on data characteristics: linear for trends, logistic for probabilities, and polynomial for complex curves. Proper pre-processing ensures accurate model outcomes.

Keywords: Model; outliers; power; simple; data; regression.

1. INTRODUCTION

When it comes to statistical analysis, a regression model is a strong tool that can be used to assess and forecast the connection that exists between a dependent variable and one or more independent variables [1]. A better understanding of how changes in the independent factors affect the dependent variable may be gained via the use of regression models, which quantify the strength and direction of the correlations between the variables. The use of this strategy is widespread across a variety of sectors, including economics, biology, engineering, and the social sciences, with the purpose of predicting events, identifying patterns, and making choices based on accurate information [2]. There are many different kinds of regression models, such as linear, logistic, and polynomial regression, and each of these modelling approaches is intended to handle a distinct category of data patterns and connections. Which model is used is determined by the characteristics of the data as well as the particular research issue that is being investigated [3]. In order to derive relevant insights and make accurate predictions, it is vital to have a solid understanding of how to effectively use and interpret regression models. In statistical analysis, one of the most important steps is selecting the appropriate regression model for the data you have. This decision will have an effect on the reliability and transferability of your conclusions [4]. The domains of data analysis, machine learning, and predictive modelling all use regression models as a fundamental component of their respective professions. Predictions, a comprehension of how variables interact with one another, and an understanding of complicated data structures are all made possible via the use of these models, which are used to investigate the links between dependent (target) and independent (predictor) variables [5].

S. Damodharan [6] studied Data-Driven Agriculture: The power of regression models deals it has revolutionized modern farming practices, enabling them to optimize their resources and maximize productivity. By examining real -world case studies, illustrate how regression models can enhance decision-making process, improve crop yields, and promote sustainable farming practices to explores the power of regression models in agriculture, discussing their applications. S. Damodharan et al. [7] studied WEKA models for rainfall data plays a vital role in India for drinking and irrigation processes. In India, there are four seasons according to seasonal adjustments. In this study, they fitted models by using a rep tree, additive regression, random subspace, and decision table using WEKA software. It gives the best estimated values, based on root absolute square error values, relative absolute error values, root relative square error, and systematic mean absolute percentage error values. S. Damodharan et al. [8] discussed the quantile regression models for rainfall data to fitted linear regression model and quantile regression model at various values of tau 0.25, 0.5, and 0.75 for Northwest India, West central India, Northeast India, Central Northeast India, and Peninsular India. Best model among fitted four models is choosing by using root mean square criteria.

1.1 What is Regression?

The statistical technique known as regression is a basic strategy that is used to investigate the

connection that exists between a single dependent variable and one or more autonomous variables. In its most fundamental form, regression analysis is concerned with gaining an understanding of how the dependent variable shifts in response to changes in any one of the independent variables, while the other independent variables remain unchanged [9]. Because it enables both prediction and inference, this method is used extensively in a variety of domains, including economics, biology, engineering, and the social sciences, among others. The simplest form of regression, known as linear regression, assumes a linear relationship between the dependent variable and the independent variable(s). This can be represented by a straight line in a two-dimensional space, where the slope of the line indicates the strength and direction of the relationship [10]. The basic equation for a simple linear regression model is:

$$Y=\beta 0+\beta 1X+\mathcal{E}$$

Where:

- Y is the dependent variable,
- X is the independent variable,
- β_0 is the intercept,
- β_1 is the slope coefficient, and
- $\mathcal{E}$ epsilon represents the error term, accounting for the variability in Y that cannot be explained by X.

When more than one independent variable is involved, the model extends to multiple linear regression, which can handle multiple predictors. The equation then becomes:

$$Y=\beta 0+\beta_1 X_1+\beta_2 X_2+...+\beta nXn+\mathcal{E}$$

Where X_1, X_2, ..., Xn are the independent variables, and β_1,β_2,...,βn are the corresponding coefficients [11].

The scope of regression analysis extends beyond the investigation of linear connections. For the purpose of capturing more complicated correlations, nonlinear regression models, such as polynomial regression or logistic regression, may be used. Polynomial regression, for instance, may be used to model curves by including higher-degree terms of the independent variable(s), while logistic regression is used for binary outcomes, modelling the chance that certain event will take place [12]. It is of the utmost is importance to choose the appropriate regression model since it is the driving force behind the precision and dependability of the predictions and inferences. A bad decision might result in skewed estimates, poor projections, and inaccurate conclusions, while a good model will capture the genuine underlying connection between the variables. Therefore, a good model will capture the true relationship. Therefore, in order to conduct an efficient data analysis, it is necessary to have a solid grasp of the assumptions, strengths, and limits of the various regression models [13].

2. IMPORTANCE OF CHOOSING THE RIGHT MODEL

Selection of the appropriate regression model is of utmost importance for a number of reasons, each of which has an effect on the quality and usefulness of the research undertaken. Choosing the right model is important for the following reasons:

1. The Properness of the Predictions

In many cases, the major objective of regression analysis is to arrive at correct predictions for data that has not yet been observed or collected. One of the reasons why an improper model might result in inaccurate predictions is that it could not accurately represent the connection that exists between the variables. One example of this would be the use of a linear model for data that is fundamentally nonlinear, which may lead to large prediction mistakes. The selection of a model that is appropriate increases the precision of predictions, which in turn makes the outcomes more trustworthy and applicable [14].

2. The Soundness of the Inferences

Inferences about the connections between variables and the testing of hypotheses are often made with the use of regression models. Any conclusions that are formed from the data may be erroneous if the model is not appropriate for the data. Assuming that there is a linear connection between two variables when there is none, for instance, might result in inaccurate inferences about the type and strength of the interactions between the variables. It is impossible to arrive at reasonable conclusions based on statistical analysis without first drawing valid inferences [15].

3. Capacity for Model Interpretation

There is a wide range of interpretability available across the various models. When conveying

findings to stakeholders or audiences who are not technically oriented, it is helpful to use simpler models, such as linear regression, since they are easier to analyse and explain. There is a possibility that more complex models, such as neural networks, have superior prediction accuracy; nevertheless, they may be difficult to understand. To choose the appropriate model, it is necessary to strike a balance between the requirements of interpretability and the requirements of forecast accuracy [16].

4. Steering clear of both overmatching and under matching

In order to avoid both overfitting and under fitting, it is important to choose the suitable model. An example of overfitting is when a model is too complicated, causing it to capture noise in the data rather than the underlying pattern. This results in poor performance when applied to fresh data examples. When a model is too simplistic to accurately represent the connection in question, a phenomenon known as under fitting occurs. This leads to insufficient performance. Selecting the appropriate model helps establish a balance, which in turn enhances the ability to generalize to new data [17].

5. Accepting the Assumptions,

In every regression model, there are certain assumptions that are made about the data itself. In the case of linear regression, for example, its assumptions include linearity, homoscedasticity, and the normality of residuals. A valid and dependable set of findings may be ensured by selecting a model that is consistent with the assumptions that have been made. If these assumptions are violated without being addressed, it is possible that the findings may be deceptive, which will also compromise the integrity of the study [18].

6. Effectiveness and numerical computation

There are a variety of computational requirements that the various models have. The computing efficiency of some models, such as linear regression, is high, but the computer resources required by other models, such as complicated machine learning techniques, are higher. In order to conduct an effective analysis and get findings in a timely manner, it is essential to choose a model that is compatible with the computing resources that are available and the size of the dataset [19].

7. The Data Characteristics Handling Process

The choice of model is influenced by the characteristics of the data, which include the distribution of the data, the existence of outliers, and the types of variables there are. The logistic regression method, for instance, is appropriate for binary outcomes, while the polynomial regression method is used to analyse nonlinear connections. When these data properties are addressed in the appropriate manner using the appropriate model, reliable analysis and improved insights are guaranteed [20].

8. The ability to reproduce and maintain consistency

When it comes to reproducibility and consistency of outcomes, selecting the appropriate model is a significant contributor. When applied to datasets that are comparable, a model that is well-suited will provide consistent findings, while an improper model may produce results that are inconsistent or untrustworthy. When it comes to confirming results and ensuring that conclusions are robust and generalizable, reproducibility is an extremely important factor [21].

9. Assurance of conformity with norms and recommended procedures

The observance of standards and best practices is absolutely necessary in a variety of sectors, including healthcare, finance, and the social sciences. By selecting the suitable model, one may assure compliance with industry standards, regulatory regulations, and scientific best practices, all of which can be very important for publishing, policy-making, and decision-making [22].

3. TYPES OF REGRESSION MODELS

1. Linear Regression

The relationship between a dependent variable (often referred to as the response or outcome variable) and one or more independent variables (often referred to as predictors or features) is modelled using linear regression, a fundamental statistical technique used in data analysis. The primary objective of linear regression is to identify the line or hyperplane that minimizes the discrepancy between the predicted values and the observed data points, depending on whether the independent variable is a single or multiple [23].

Key Concepts of Linear Regression:

Simple Linear Regression:

Model: $y = \beta_0 + \beta_1 x + \varepsilon$; y = β_0 + β_1 X + Ɛ

The dependent variable is y y, the independent variable is x x, the y-intercept is β 0 β 0, the slope of the line is β 1 β 1, and the error term is ϵ ϵ.
The objective is to minimize the sum of the squared differences between the predicted and observed values by estimating the parameters β_0 β_0 and β_1 β_1 [24].

Multiple linear regression:

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$; $y=\beta_0 +\beta_1x_1+\beta_2x_2+\cdots+\beta_n x_n+ \varepsilon$
In this model, the dependent variable is still y, but there are now multiple independent variables x_1, x_2, x_3, x_4, x_5, and x_6. Each variable has its own coefficient, β_1, β_2, β_3, β_4, β_5, and β_6.

The objective is to estimate the parameters in a manner that ensures the predicted y y values are as close as possible to the actual y y values across multiple dimensions, while maintaining the same principles as simple linear regression [25].

Linear Regression Assumptions

- Linearity: The dependent and independent variables should exhibit a linear relationship.
- Independence: It is essential that observations are not dependent on one another.
- Homoscedasticity: The variance of residuals (errors) should remain constant at all levels of the independent variables.
- Normality: The residuals of the model should be approximately normally distributed [26].

Fitting a Linear Regression Model:

Ordinary Least Squares (OLS) is the most prevalent method for fitting a linear regression model. This method minimizes the sum of the squared residuals, which are the discrepancies between the predicted and observed values.

R-squared: A statistical measure that quantifies the extent to which the independent variables predict the variance of the dependent variable. It is variable, ranging from 0 to 1 [27].

Utilizations:

- Predictive modelling is the process of predicting future values by analysing historical data. Identifying trends in data over time [28].
- Risk Assessment: The assessment of risk factors in the fields of finance, insurance, and other relevant areas.
- Market Research: Comprehending the influence of a variety of factors on consumer behaviour.

For instance, suppose that an organization desires to forecast its revenues by analysing its advertising expenditures. They accumulate information regarding previous advertising expenditures (in thousands of dollars) and their corresponding sales (in thousands of units). They have the option of employing linear regression to determine the correlation between advertising expenditures and sales. If the final model is the following: Sales = 2 + 0.5 × Advertising, it indicates that:

Sales increase by 0.5 thousand units (500 units) for every $1,000 spent on advertising. The base sales level is 2,000 units when there is no advertising expenditure [29]. Simple Linear Regression: A scatter plot that depicts the best-fit line as a straight line [30]. Residual depiction: A depiction of residuals that is used to verify the assumptions of linear regression. Linear regression is an extensively used method for modelling linear relationships in a variety of disciplines, including economics, engineering, and social sciences and due to its simplicity, interpretability and effectiveness [31].

2. Polynomial Regression

A polynomial regression is an extension of linear regression that represents the relationship between the independent variable(s) and the dependent variable as an n-degree polynomial. Polynomial regression is capable of fitting data points with a curvilinear or more complex pattern, in contrast to linear regression, which implies a straight-line relationship [32].

The fundamental structure of a polynomial regression model is as follows:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$; $y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_n x_n+ \varepsilon$

In this context, the dependent variable is y y, the independent variable is x x, the coefficients are β 0 , β 1 , … , βn β 0 , β 1 , …, β n, the degree of the polynomial is n n, and the error term is Ɛ [33,34].

The independent variable x is raised to higher powers in polynomial regression to model more complex relationships. For instance,

A parabolic relationship is represented by a quadratic regression (degree 2): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$; $y=\beta_0 + \beta_1 x_1+\beta_2 x_2$

.

An S-shaped curve can be represented by a cubic regression (degree 3) as follows:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$; $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$ [35].

What is the rationale for employing polynomial regression?

Non-Linear Relationships: Polynomial regression can offer a superior approximation than linear regression when the relationship between the independent and dependent variables is non-linear [36].

Flexibility: It enables the modelling of data with curves, which is not possible with a straight line (linear regression) [37].

Fitting a Polynomial Regression Model:

Data Transformation: Polynomial regression is a process that entails the generation of new features by elevating the independent variable to the powers of the desired degree. For example, the quadratic regression model comprises two features: x x and x 2 x 2.

Least Squares Method: Polynomial regression, similar to linear regression, typically employs the least squares method to estimate the coefficients β_0 , β_1, … , β_n ; $\beta_0,\beta_1,\ldots,\beta n$ [38,39,40].

Overfitting:

Risk of Overfitting: Overfitting is a phenomenon in which a model fits the training data exceptionally well but performs unfavourably on new data. This can occur when higher-degree polynomials are employed. This is due to the possibility that the model may capture disturbance in the data rather than the underlying pattern [41,42].

Model Selection: It is crucial to exercise caution when selecting the polynomial degree, frequently employing cross-validation to prevent overfitting [43,44].

Visualization:

The fitted polynomial curve can be plotted in conjunction with the data points to help visualize polynomial regression. The complexity of the trajectory is determined by the degree of the polynomial [45].

For instance,

Assume that you possess information regarding the halting distance and velocity of a vehicle. The relationship between speed and halting distance is not linear; the stopping distance is not merely doubled when the speed is doubled; it is increased quadratically [46]. A quadratic regression model may be suitable in this scenario:

Stopping Distance = β_0 + β_1 × Speed + β_2 × Speed + $\beta 2$; Stopping Distance = β 0 + β 1 × Speed + β 2 × Speed + $\beta 2$

This model will more effectively represent the non-linear relationship than a simple linear model [47].

Utilizations:

Economics: The simulation of non-linear relationships between economic indicators, such as the influence of income on consumption [48].

Medicine: The simulation of dose-response curves.

Engineering: The modelling of the stress-strain relationship in materials, which may be quadratic or cubic [49].

Visualization: In contrast to linear regression, which involves a straight line, polynomial regression involves a fitted curve that follows the trend of the data points in a seamless manner. The model's ability to suit the data across various ranges of the independent variable is demonstrated by a scatter diagram with the polynomial curve [50].

Although polynomial regression is effective in identifying intricate patterns in data, it must be implemented with caution to prevent overfitting and guarantee the generalizability of the model. A polynomial regression is an extension of linear regression that represents the relationship between the independent variable(s) and the dependent variable as an n-degree polynomial. Polynomial regression is capable of fitting data points with a curvilinear or more complex pattern, in contrast to linear regression, which implies a straight-line relationship [51,52].

3. Logistic Regression

A categorical dependent variable's outcome is predicted using logistic regression, a form of regression analysis that is based on one or more independent variables. This method is frequently employed when the dependent variable is binary, meaning that it has two potential outcomes, such as "yes" or "no," "success" or "failure," etc.

1. Binary Dependent Variable: -The dependent variable in logistic regression is typically binary (e.g., True or False, 0 or 1) [53].

- Predicting whether a student will pass or fail an exam based on the number of hours spent studying, for instance [54].

2. Sigmoid Function: - The logistic function, also known as the sigmoid function, is employed in logistic regression to approximate the likelihood that a provided input \(x\) is a member of a specific class [55].

- The definition of the sigmoid function is as follows: σ(z)=1+e−z1

- The linear combination of the input variables, \ (z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n \), is mapped to a value between 0 and 1, which represents the probability of the dependent variable being in one class, by \(\sigma(z) \).

3. Model:

- The logistic regression model is depicted as: \[P(y=1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1x_1 + \dots + \beta_nx_n)}} \]

- The probability that the dependent variable \(y \) equals 1 (e.g., success) in the presence of the input \(x \) is denoted by \(P(y=1|x) .

- The coefficients \(\beta_0, \beta_1, \dots, \beta_n \) are estimated by the model using a method known as Maximum Likelihood Estimation (MLE), which identifies the parameters that provide the most reliable explanation for the observed data [56,57].

4. Interpretation: - The coefficients \(\beta_i \) denote the change in the log-odds of the dependent variable for a one-unit change in the corresponding independent variable \(x_i).

A prediction can be made by converting the log-odds to odds and then to probabilities [58]. 5. Decision Boundary:- The threshold probability at which the model predicts one class or the other is referred to as the decision boundary [59]. For binary outcomes, this threshold is frequently established at 0.5, which translates to:

- Assume that \(P(y=1|x) \geq 0.5 \) and \(y = 1 \).
- Predict that y is equal to zero if the probability of y being equal to one when x is equal to one is less than 50%.

6. Assumptions: - Independence: Observations should be separate from one another.
- Linearity of Log-Odds: The independent variables, as well as the log-odds of the dependent variable, should be linearly related.

- Independent variables should not be strongly correlated: This prevents multicollinearity. 7. Extensions: - Multinomial Logistic Regression: Employed when the dependent variable has more than two categories (e.g., predicting the type of fruit: apple, banana, orange).
-- Ordinal Logistic Regression: Employed when the dependent variable is ordinal (i.e., the categories have a natural order, such as low, medium, and high).

8. Metrics for Evaluation: - Accuracy: The percentage of accurate predictions.

- Precision and Recall: Applications for datasets that are imbalanced, with one class being more prevalent than the other.

- ROC Curve and AUC: The ROC curve illustrates the true positive rate in relation to the false positive rate, while the AUC (Area under the Curve) assesses the model's capacity to differentiate between classes.

Here is an example [60,61,62]:

We will assume that you wish to forecast whether an individual will purchase a product (yes or no) based on their income and age. A logistic regression model could be constructed as follows: \[P(\text{Buy} = 1|\text{Age}, \text{Income}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Income})}}.

An output probability between 0 and 1 is generated by the model. If the probability exceeds 0.5, the model anticipates that the individual will purchase the product; otherwise, it anticipates that they will not [63].

Uses:

- Medical Research: Analysing patient data to predict the presence or absence of a disease.

- Marketing: Making predictions regarding the responsiveness of a consumer to a marketing campaign.

- Finance: Calculating the probability of a customer defaulting on a loan through credit assessment.

-- Social Sciences: The process of analysing survey data to ascertain the factors that influence a specific behaviour [64,65].

Visualization:

In a logistic regression model with a single independent variable, the decision boundary can be displayed on a scatter plot, with the sigmoid curve indicating the probability of the outcome as the independent variable varies.

In a variety of disciplines, including finance, medicine, marketing, and social sciences, logistic regression is frequently employed due to its simplicity, interpretability, and effectiveness in binary classification tasks [66].

4. KEY CONSIDERATIONS FOR MODEL SELECTION

Selecting the appropriate model for a given task is a critical step in the data analysis process. The choice of model impacts not only the performance but also the interpretability, complexity, and robustness of the results [67,68,69]. Here are key considerations for model selection:

1. Nature of the Problem:

- **Type of Prediction**: Determine whether the problem is one of regression (predicting continuous outcomes) or classification (predicting categorical outcomes).
- **Outcome Variable**: Consider the type of the dependent variable (binary, continuous, multinomial) to guide model selection.

2. Interpretability:

- **Model Transparency**: Some models, like linear regression or decision trees, are more interpretable, meaning it's easier to understand how the input features are influencing the output.
- **Regulatory Requirements**: In some fields, such as healthcare or finance, models need to be interpretable to meet regulatory standards.

3. Model Complexity:

- **Overfitting vs. under fitting**: Complex models like deep neural networks can over fit, capturing noise in the training data, while simple models might under fit, failing to capture underlying patterns.
- **Bias-Variance Trade-off**: Balancing model complexity to achieve a trade-off between bias (error due to overly simplistic models) and variance (error due to overly complex models).

4. Data Characteristics:

- **Size of the Dataset**: Large datasets might support more complex models like deep learning, whereas smaller datasets might require simpler models to avoid overfitting.
- **Feature Set**: The number and nature of features (e.g., numeric, categorical, missing values) can influence the choice of model. For instance, models like XGBoost handle missing data natively.
- **Feature Distribution**: Consider whether the data follows a normal distribution, is skewed, or contains outliers, as this can affect the performance of certain models.

5. Computational Efficiency:

- **Training Time**: Some models, like deep learning models, require significant

computational resources and time to train, especially on large datasets.
- **Scalability**: Consider how well the model scales with increasing data size or feature dimensionality.
- **Prediction Speed**: In real-time applications, the speed at which the model can make predictions is critical.

6. Model Performance:

- **Accuracy**: Evaluate the model's predictive accuracy using metrics appropriate for the task, such as RMSE for regression or accuracy, precision, recall, F1-score, and AUC for classification.
- **Cross-Validation**: Use cross-validation to assess the model's performance on unseen data and to avoid overfitting.
- **Robustness**: Assess the model's ability to perform well on slightly different data distributions, checking for robustness to noise and outliers.

7. Regularization:

- **Preventing Overfitting**: Regularization techniques like Ridge, Lasso, or Elastic Net can be important in controlling model complexity and improving generalization.
- **Hyper parameter Tuning**: Regularized models often require careful tuning of hyper parameters to balance the model's fit to the data and its complexity.

8. Availability of Domain Knowledge:

- **Incorporating Expertise**: Models like Bayesian networks or rule-based systems can incorporate domain knowledge, which might be crucial in some applications.
- **Interpretation in Context**: How well the model's predictions can be interpreted in the context of the specific domain can influence model choice.

9. Model Stability:

- **Sensitivity to Data Changes**: Assess how sensitive the model is to small changes in the input data. Models that are too sensitive may not generalize well to new data.
- **Reproducibility**: Ensure that the model produces consistent results when trained on different samples from the same dataset.

10. Deployment Considerations:

- **Production Environment**: Consider the environment where the model will be deployed. Some models might require more computational resources or specific software that might not be available in all deployment environments.
- **Maintainability**: Complex models might be harder to maintain and update over time compared to simpler models.

11. Cost of Errors:

- **Error Implications**: In some applications, the cost of false positives might be higher than false negatives (or vice versa), influencing the choice of model and the metrics used to evaluate it.
- **Risk Management**: Consider models that can provide a measure of uncertainty in their predictions if the application requires managing risk.

12. Availability of Tools and Expertise:

- **Tool Availability**: Ensure that the tools and libraries required for the model are available and well-supported.
- **Expertise**: Choose a model that aligns with the team's expertise to ensure it can be properly implemented, interpreted, and maintained.

5. COMMON PITFALLS IN REGRESSION ANALYSIS

Selecting the appropriate model for a given task is a critical step in the data analysis process. The choice of model impacts not only the performance but also the interpretability, complexity, and robustness of the results. Here are key considerations for model selection:

1. Nature of the Problem:

- **Type of Prediction**: Determine whether the problem is one of regression (predicting continuous outcomes) or classification (predicting categorical outcomes).
- **Outcome Variable**: Consider the type of the dependent variable (binary, continuous, multinomial) to guide model selection.

2. Interpretability:

- **Model Transparency**: Some models, like linear regression or decision trees, are

more interpretable, meaning it's easier to understand how the input features are influencing the output.
- **Regulatory Requirements**: In some fields, such as healthcare or finance, models need to be interpretable to meet regulatory standards.

3. Model Complexity:

- **Overfitting vs. Underfitting**: Complex models like deep neural networks can overfit, capturing noise in the training data, while simple models might underfit, failing to capture underlying patterns.
- **Bias-Variance Trade-off**: Balancing model complexity to achieve a trade-off between bias (error due to overly simplistic models) and variance (error due to overly complex models).

4. Data Characteristics:

- **Size of the Dataset**: Large datasets might support more complex models like deep learning, whereas smaller datasets might require simpler models to avoid overfitting.
- **Feature Set**: The number and nature of features (e.g., numeric, categorical, missing values) can influence the choice of model. For instance, models like XGBoost handle missing data natively.
- **Feature Distribution**: Consider whether the data follows a normal distribution, is skewed, or contains outliers, as this can affect the performance of certain models.

5. Computational Efficiency:

- **Training Time**: Some models, like deep learning models, require significant computational resources and time to train, especially on large datasets.
- **Scalability**: Consider how well the model scales with increasing data size or feature dimensionality.
- **Prediction Speed**: In real-time applications, the speed at which the model can make predictions is critical.

6. Model Performance:

- **Accuracy**: Evaluate the model's predictive accuracy using metrics appropriate for the task, such as RMSE for regression or accuracy, precision, recall, F1-score, and AUC for classification.
- **Cross-Validation**: Use cross-validation to assess the model's performance on unseen data and to avoid overfitting.
- **Robustness**: Assess the model's ability to perform well on slightly different data distributions, checking for robustness to noise and outliers.

7. Regularization:

- **Preventing Overfitting**: Regularization techniques like Ridge, Lasso, or Elastic Net can be important in controlling model complexity and improving generalization.
- **Hyperparameter Tuning**: Regularized models often require careful tuning of hyperparameters to balance the model's fit to the data and its complexity.

8. Availability of Domain Knowledge:

- **Incorporating Expertise**: Models like Bayesian networks or rule-based systems can incorporate domain knowledge, which might be crucial in some applications.
- **Interpretation in Context**: How well the model's predictions can be interpreted in the context of the specific domain can influence model choice.

9. Model Stability:

- **Sensitivity to Data Changes**: Assess how sensitive the model is to small changes in the input data. Models that are too sensitive may not generalize well to new data.
- **Reproducibility**: Ensure that the model produces consistent results when trained on different samples from the same dataset.

10. Deployment Considerations:

- **Production Environment**: Consider the environment where the model will be deployed. Some models might require more computational resources or specific software that might not be available in all deployment environments.
- **Maintainability**: Complex models might be harder to maintain and update over time compared to simpler models.

11. Cost of Errors:

- **Error Implications**: In some applications, the cost of false positives might be higher

than false negatives (or vice versa), influencing the choice of model and the metrics used to evaluate it.
- **Risk Management**: Consider models that can provide a measure of uncertainty in their predictions if the application requires managing risk.

12. Availability of Tools and Expertise:

- **Tool Availability**: Ensure that the tools and libraries required for the model are available and well-supported.
- **Expertise**: Choose a model that aligns with the team's expertise to ensure it can be properly implemented, interpreted, and maintained.

6. COMMON PITFALLS IN REGRESSION ANALYSIS

Regression analysis is a powerful tool for understanding relationships between variables and making predictions. However, several common pitfalls can undermine the reliability of regression models and lead to misleading conclusions [70,71,72]. Here are some common pitfalls in regression analysis and how to address them:

1. Ignoring Assumptions

Assumptions in Linear Regression:

- **Linearity**: The relationship between independent and dependent variables should be linear.
- **Independence**: Observations should be independent of each other.
- **Homoscedasticity**: The variance of residuals should be constant across levels of the independent variables.
- **Normality of Residuals**: Residuals should be approximately normally distributed.

Pitfall: Violating these assumptions can lead to biased estimates and incorrect inferences.
Solution: Check and validate assumptions using diagnostic plots (e.g., residuals vs. fitted values) and statistical tests. If assumptions are violated, consider transformations or alternative models.

2. Multicollinearity

Description: Multicollinearity occurs when independent variables are highly correlated with each other, making it difficult to isolate the effect of each predictor on the dependent variable.
Pitfall: Multicollinearity can lead to inflated standard errors, making it harder to determine the significance of predictors.
Solution: Diagnose multicollinearity using variance inflation factors (VIFs). Address it by removing or combining correlated variables, or using techniques like Ridge Regression that handle multicollinearity.

3. Overfitting

Description: Overfitting happens when a model is too complex and captures noise in the training data rather than the underlying pattern.
Pitfall: An over fitted model will perform well on training data but poorly on unseen data.
Solution: Use techniques such as cross-validation to assess model performance on new data. Regularization methods like Ridge and Lasso can help prevent overfitting.

4. Under fitting

Description: Under fitting occurs when a model is too simple to capture the underlying pattern in the data.
Pitfall: An under fitted model will have high bias and poor performance on both training and test data.
Solution: Use more complex models or add interaction terms and polynomial features if appropriate. Evaluate model performance using metrics and adjust the model as needed.

5. Omitted Variable Bias

Description: This bias occurs when a relevant variable is omitted from the model, leading to incorrect estimates of the coefficients for included variables.
Pitfall: Omitted variable bias can distort the estimated relationships between variables.
Solution: Carefully select and include all relevant variables based on theory and prior research. Conduct sensitivity analyses to check the impact of omitted variables.

6. Data Snooping

Description: Data snooping (or p-hacking) involves multiple testing or reusing the same data for model selection and hypothesis testing, which can inflate Type I error rates.
Pitfall: Data snooping can lead to misleading conclusions and overestimated model performance.
Solution: Use proper statistical techniques and reserve a separate dataset or cross-validation

approach for model testing. Avoid multiple testing without correction.

7. Endogeneity

Description: Endogeneity arises when an independent variable is correlated with the error term, often due to omitted variables, measurement error, or simultaneity.
Pitfall: Endogeneity can lead to biased and inconsistent estimates.
Solution: Use instrumental variable (IV) techniques or other methods to address endogeneity. Ensure that chosen instruments are valid and relevant.

8. Outliers and Influential Points

Description: Outliers or influential data points can disproportionately affect the results of a regression analysis.
Pitfall: Outliers can skew results and lead to incorrect conclusions about the relationships between variables.
Solution: Identify and analyse outliers using diagnostic tools like leverage and Cook's distance. Consider robust regression methods if outliers are present.

9. Incorrect Functional Form

Description: Using an incorrect model form (e.g., linear when the relationship is nonlinear) can lead to misleading results.
Pitfall: An incorrect functional form can miss important relationships and lead to poor model fit.
Solution: Explore different functional forms and transformations of variables. Use residual plots and other diagnostic tools to check for model fit.

10. Sample Size Issues

Description: Small sample sizes can lead to unreliable estimates and increased variability in the results.
Pitfall: Small sample sizes may not provide enough information to accurately estimate the model parameters and their significance.
Solution: Ensure an adequate sample size for the complexity of the model. Use techniques like bootstrapping for more robust estimates if sample size is a concern.

11. Misinterpretation of Results

Description: Misinterpreting the output of a regression model, such as misunderstanding the meaning of coefficients or significance levels.
Pitfall: Misinterpretation can lead to incorrect conclusions about the relationships between variables.
Solution: Carefully interpret regression coefficients in the context of the model. Use appropriate metrics and confidence intervals to assess the reliability of the estimates.

12. Ignoring Model Validation

Description: Failing to validate the model on new or unseen data can lead to overly optimistic performance estimates.
Pitfall: Without proper validation, it's difficult to gauge how well the model will generalize to new data.
Solution: Use techniques such as cross-validation or hold-out validation to evaluate model performance on independent data sets.

7. TOOLS AND TECHNIQUES FOR MODEL SELECTION AND EVALUATION

Model selection and evaluation are crucial steps in building effective predictive models. A variety of tools and techniques are available to help you choose the best model and assess its performance [73]. Here's a comprehensive guide to some of the most commonly used methods:

1. Model Selection Techniques

1. Cross-Validation:

- **Purpose**: To assess how the results of a statistical analysis generalize to an independent data set.
- **Methods**:
  - **K-Fold Cross-Validation**: Divides the dataset into kkk subsets, trains on k−1k-1k−1 subsets, and tests on the remaining subset. Repeats for each subset.
  - **Leave-One-Out Cross-Validation (LOOCV)**: A special case of k-fold where kkk is equal to the number of data points. Each data point is used as a test set once.
- **Tools**: Scikit-learn's cross_val_score, GridSearchCV, RandomizedSearchCV.

2. Grid Search:

- **Purpose**: To systematically work through multiple combinations of parameter values, cross-validating as it goes to determine which combination gives the best performance.
- **Tools**: Scikit-learn's GridSearchCV.

3. Random Search:

- **Purpose**: To randomly sample from a range of hyperparameter values rather than testing all possible combinations.
- **Tools**: Scikit-learn's Randomized Search CV.

4. Model Comparison:

- **Purpose**: To compare the performance of different models using the same data.
- **Tools**: Comparison of metrics like accuracy, precision, recall, F1 score, ROC AUC, etc., using libraries like Scikit-learn, Statsmodels, or custom scripts.

2. Evaluation Metrics

1. Regression Metrics:

- **Mean Absolute Error (MAE)**: Average absolute difference between actual and predicted values.
- **Mean Squared Error (MSE)**: Average of the squares of the errors.
- **Root Mean Squared Error (RMSE)**: Square root of MSE; provides error in the same unit as the target variable.
- **R-Squared (R^2)**: Proportion of variance in the dependent variable that is predictable from the independent variables.

2. Classification Metrics:

- **Accuracy**: Proportion of correctly classified instances out of the total instances.
- **Precision**: Proportion of true positives among predicted positives.
- **Recall (Sensitivity)**: Proportion of true positives among actual positives.
- **F1 Score**: Harmonic mean of precision and recall.
- **ROC Curve and AUC**: ROC curve plots true positive rate vs. false positive rate. AUC is the area under this curve.
- **Confusion Matrix**: A table layout that visualizes performance of a classification algorithm.

3. Clustering Metrics:

- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters.
- **Davies-Bouldin Index**: Measures the average similarity ratio of each cluster with its most similar one.

3. Model Validation

1. Train-Test Split:

- **Purpose**: To divide the dataset into training and testing sets to evaluate how well the model generalizes to unseen data.
- **Tools**: Scikit-learn's train_test_split.

2. Bootstrap:

- **Purpose**: To estimate the accuracy of a model by repeatedly resampling the data with replacement and evaluating performance.
- **Tools**: Scikit-learn's Bootstrap.

3. Resampling Methods:

- **Purpose**: To estimate the distribution of a statistic by resampling the data.
- **Types**: Bootstrapping, Jackknife.

4. Feature Selection and Importance

1. Recursive Feature Elimination (RFE):

- **Purpose**: To recursively remove features and build a model on the remaining features to identify the most important ones.
- **Tools**: Scikit-learn's RFE.

2. Feature Importance:

- **Purpose**: To assess the importance of each feature in the model's predictions.
- **Tools**: Feature importance attributes of models like Random Forest, XGBoost.

3. Principal Component Analysis (PCA):

- **Purpose**: To reduce dimensionality by transforming features into a set of linearly uncorrelated components.
- **Tools**: Scikit-learn's PCA.

5. Hyper parameter Tuning

1. Bayesian Optimization:

- **Purpose**: To use a probabilistic model to optimize hyperparameters.
- **Tools**: Libraries like hyperopt, Optuna.

2. Genetic Algorithms:

- **Purpose**: To use evolutionary techniques to search for optimal hyperparameters.
- **Tools**: Libraries like TPOT.

6. Model Robustness and Stability

1. Sensitivity Analysis:

- **Purpose**: To assess how sensitive the model's predictions are to changes in input values or parameters.

2. Robustness Checks:

- **Purpose**: To check model stability under different conditions or with different subsets of the data.

7. Visualization Tools

1. Residual Plots:

- **Purpose**: To visualize residuals to check for patterns indicating model misfit.
- **Tools**: Matplotlib, Seaborn.

2. Learning Curves:

- **Purpose**: To plot model performance as a function of training size or epochs.
- **Tools**: Scikit-learn's learning_curve.

3. ROC Curves:

- **Purpose**: To visualize the trade-off between true positive rate and false positive rate.
- **Tools**: Scikit-learn's roc_curve, roc_auc_score.

4. Feature Importance Plots:

- **Purpose**: To visualize the relative importance of features.
- **Tools**: Matplotlib, Seaborn.

8. CONCLUSION

A thorough familiarity with the data, the study subject, and the assumptions supporting various models is necessary for choosing the right regression model. A number of recommended practices should be adhered to in order to guarantee dependable and strong model selection. To begin, use visualizations and summary statistics to your advantage in an Exploratory Data Analysis (EDA) to fully grasp the data's properties. In order to find problems, trends, and correlations in the dataset, this first stage is essential. It is equally important to do diagnostic checks, which include testing and plotting to validate model assumptions and make sure the selected model fits the data well. To ensure the model generalizes effectively to new data, it is crucial to use cross-validation to assess its performance and reduce the likelihood of overfitting. When working with several predictors or dealing with multicollinearity, it is recommended to use regularization approaches to enhance the resilience and performance of the model. Also, keep things basic; simpler models are easier to understand and use, and they may frequently explain plenty without being too complicated. Finally, as regression analysis develops, it is essential to continue learning; doing so will allow you to make better judgments and use best practices when selecting and evaluating models.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of this manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ahn H, Loh W. Tree-structured proportional hazards regression modeling. Biometrics. 1994;50:471–485.
2. Altman DG. Categorising continuous covariates (letter to the editor). Brit J Cancer. 1991;64:975.
3. Altman DG. Suboptimal analysis using 'optimal' cutpoints. Brit J Cancer. 1998;78: 556–557.
4. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. J Nat Cancer Inst. 1994;86:829–835.
5. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med. 2007;26:2937–2957.
6. Damodharan S. Data – Driven Agriculture: The power of regression models, The Agriculture Magazine, vol-3, issue-10, E-ISSN: 2583-1755, PP.475-479.
7. Damodharan S, Venkataramana Reddy S, Sarojamma B. WEKA Models for rainfall

data, Journal of Emerging Technologies and Innovative Research (JETIR), vol.9, issue. 9, ISSN- 2349-5162,
DOI- JETIR2209232, PP: C11 – C16.

8. Damodharan S, Venkataramana Reddy S, Sarojamma B. Quantile regression models for rainfall data. International Journal of Computer Sciences and Engineering, Vol. 9, Issue. 9, ISSN: 2347-2693, PP: 83-85.
9. Belcher H. The concept of residual confounding in regression models and some applications. Stat Med. 1992;11: 1747–1758.
10. Berhane K, Hauptmann M, Langholz B. Using tensor product splines in modeling exposure–time–response relationships: Application to the Colorado Plateau Uranium Miners cohort. Stat Med. 2008;27: 5484–5496.
11. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth and Brooks/Cole, Pacific Grove, CA; 1984.
12. Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma. J Clin Epi. 1997;50: 1201–1210.
13. Chambers JM, Hastie TJ. Editors. Statistical Models in S. Wadsworth and Brooks/Cole, Pacific Grove, CA; 1992.
14. Ciampi A, Negassa, Lou Z. Tree-structured prediction for censored survival data and the Cox model. J Clin Epi. 1995;48:675–689.
15. Ciampi J, Thiffault, Nakache JP, Asselain B. Stratification by stepwise regression, correspondence analysis and recursive partition. Comp Stat Data Analysis. 1986; 185–204.
16. Clark LA, Pregibon D. Tree-Based Models. In Chambers JM, Hastie TJ. Editors, Statistical Models in S, chapter 9, pages 377–419. Wadsworth and Brooks/Cole, Pacific Grove, CA; 1992.
17. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 1979;74:829–836.
18. Cook EF, Goldman L. Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies. Am J Epi. 1988;127:626–639.
19. Cox DR. The regression analysis of binary sequences (with discussion). J Roy Stat Soc B. 1958;20:215–242.
20. Cox DR. Regression models and life-tables (with discussion). J Roy Stat Soc B. 1972;34:187–220.
21. Crichton NJ, Hinde JP, Marchini J. Models for diagnosing chest pain: Is CART useful? Stat Med. 1997;16:717–727.
22. Davis RB, Anderson JR. Exponential survival trees. Stat Med. 1989;8:947–961.
23. De Boor. A Practical Guide to Splines. Springer-Verlag, New York, revised edition; 2001.
24. Devlin TF, Weeks BJ. Spline functions for logistic regression modeling. In Proceedings of the Eleventh Annual SAS Users Group International Conference, pages 646–651, Cary, NC, SAS Institute, Inc; 1986.
25. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med. 1989; 8:551–561.
26. Faraggi, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. Stat Med. 1996;15:2203–2213.
27. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. Pharm Stat. 2009;8:50–61.
28. Friedman JH. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics. Department of Statistics, Stanford University; 1984.
29. Giannoni R, Baruah, Leong T, Rehman MB, Pastormerlo LE, Harrell FE, Coats AJ, Francis DP. Do optimal prognostic thresholds in continuous physiological variables really exist? Analysis of origin of apparent thresholds, with systematic review for peak oxygen consumption, ejection fraction and BNP. Plos One. 2014;9(1).
30. Govindarajulu US, Spiegelman D, Thurston SW, Ganguli B, Eisen EA. Comparing smoothing techniques in Cox models for exposure-response relationships. Stat Med. 2007;26:3735–3752.
31. Grambsch PM, O'Brien PC. The effects of transformations and preliminary tests for non-linearity in regression. Stat Med. 1991; 10:697–709.
32. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc. 1992;87:942–951.
33. Gray RJ. Spline-based tests in survival analysis. Biometrics. 1994;50:640–652.

34. Gustafson P. Bayesian regression modeling with interactions and smooth effects. J Am Stat Assoc. 2000;95:795–806.
35. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. Ca Trt Rep. 1985;69: 1071–1077.
36. Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: Determining relationships between predictors and response. J Nat Cancer Inst. 1988;80:1198–1202.
37. Hastie T. Discussion of the use of polynomial splines and their tensor products in multivariate function estimation" by C. J. Stone. Appl Stat. 1994; 22:177–179.
38. Hastie T, Tibshirani R. Generalized Additive Models. Chapman and Hall, London; 1990.
39. Hilsenbeck SG, Clark GM. Practical p-value adjustment for optimally selected cutpoints. Stat Med. 1996;15: 103–112.
40. Holländer N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. Stat Med. 2004;23: 1701–1713.
41. Keleş S, Segal MR. Residual-based tree-structured survival analysis. Stat Med. 2002;21:313–326.
42. Lausen B, Schumacher M. Evaluating the effect of optimized cut off values in the assessment of prognostic factors. Comp Stat Data Analysis. 1996; 21(3):307–326.
43. LeBlanc M, Crowley J. Survival trees by goodness of fit. J Am Stat Assoc. 1993; 88:457–467.
44. Magee L. Nonlocal behavior in polynomial regressions. Am Statistician. 1998;52:20–22.
45. Marshall RJ. The use of classification and regression trees in clinical epidemiology. J Clin Epi. 2001;54:603–609.
46. Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. Psych Bull. 1993l;113:181–190.
47. McNeil R, Trussell J, Turner JC. Spline interpolation of demographic data. Demography. 1977;14:245–252.
48. Moser BK, Coombs LP. Odds ratios for a continuous outcome variable without dichotomizing. Stat Med. 2004;23:1843–1860.
49. Ragland R. Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. Epi, 3:434–440. See letters to editor. 1993; 274-1992;4(3).
50. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Appl Stat. Discussion. 1994;43: 453–467.
51. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127–141.
52. Schemper M. Non-parametric analysis of treatment-covariate interaction in the presence of censoring. Stat Med. 1988;7:1257–1266.
53. Schmoor C, Ulm K, Schumacher M. Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. Stat Med. 1993; 12:2351–2366.
54. Schulgen B, Lausen J, Olsen, Schumacher M. Outcome-oriented cutpoints in quantitative exposure. Am J Epi. 1994; 120:172–184.
55. Segal MR. Regression trees for censored data. Biometrics. 1988;44:35–47.
56. Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. J Am Stat Assoc. 1990;85:941–949.
57. Smith PL. Splines as a useful and convenient statistical tool. Am Statistician. 1979;33:57–62.
58. Stone CJ. Comment: Generalized additive models. Statistical Sci. 1986;1:312–314.
59. Stone CJ, Koo CY. Additive splines in statistics. In Proceedings of the Statistical Computing Section ASA, pages 45–48, Washington, DC; 1985.
60. Suissa S, Blais L. Binary regression with continuous outcomes. Stat Med. 1995;14: 247–255.
61. Van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology. 2014;14 (1):137.
62. Wainer. Finding what is not there through the unfortunate binning of results: The Mendel effect. Chance. 2006;19(1):49–56.

63. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika. 1967;54:167–178.
64. Walter R, Feinstein AR, Wells CK. Coding ordinal independent variables in multiple regression analyses. Am J Epi. 1987;125: 319–323.
65. Wang Y, Wahba G, Gu C, Klein R, Klein B. Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. Stat Med. 1997;16:1357–1376.
66. Zhang. Classification trees for multiple binary responses. J Am Stat Assoc. 1998; 93:180–193.
67. Zhang T. Holford, Bracken MB. A tree-based method of analysis for prospective studies. Stat Med. 1996;15: 37–49.
68. Denuit M, Lang S. Nonlife ratemaking with Bayesian GAM's. Insurance: Mathematics and Economics. 2005;35:627–647.
69. Fahrmeir L, Kneib T. Bayesian smoothing and regression for longitudinal, spatial and event history data. Oxford: Oxford University Press; 2011.
70. Kneib T, Fahrmeir L. A mixed model approach for geoadditive hazard regression. Scandinavian Journal of Statistics. 2007;34:207–228.
71. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. Applied Statistics. 2005;54:507–554.
72. Rigby RA, Stasinopoulos DM. A flexible regression approach using GAMLSS in R; 2009.
Available at http://gamlss.org/.
73. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer; 2000.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/123452