International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 09-16
© 2024 Stats & Maths
https://www.mathsjournal.com
Received: 08-02-2024
Accepted: 11-03-2024

Kanwal Preet Singh Attwal
Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

Amardeep Singh Dhiman
Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

Corresponding Author:
Kanwal Preet Singh Attwal
Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

Exploring SPSS statistics for data mining and statistical modeling

Kanwal Preet Singh Attwal and Amardeep Singh Dhiman

DOI: https://doi.org/10.22271/maths.2024.v9.i3a.1721

Abstract
In the era of big data, where an astonishing volume of information, measured in billions of bytes, is generated daily, the role of data mining tools becomes paramount in extracting valuable insights. This study provides an in-depth exploration of SPSS Statistics, a powerful data mining tool that assumes a central role in data manipulation, analysis, and presentation. The paper illuminates SPSS's fundamental interface, focusing on its core components: the Data Editor and the Viewer. Within the Data Editor, we delve into its dual perspectives-the Data View and the Variable View-examining their functionalities comprehensively. Special emphasis is placed on data entry procedures within SPSS, as well as the versatile columns available in Variable View for defining variable characteristics. Furthermore, this paper elucidates the concept of statistical modeling and its practical application in constructing Regression models using SPSS. Various methods for developing robust Regression models, including Hierarchical and Stepwise Regression, are outlined. In the final section, the paper delves into diverse statistical concepts, such as the null hypothesis and significance level, offering a comprehensive discussion.

Keywords: SPSS statistics, data editor, data mining, data

Introduction
A study has been conducted to develop a model for prediction of wheat yield in Patiala district of Punjab, India [1]. As part of this study, a framework has already been proposed to develop a model for crop yield prediction [2]. A study has been conducted to show how to develop a classification model using Data Mining tool - WEKA [3]. The current paper provides an in-depth insight of SPSS Statistics - exploring its interface and discussing how to create a Regression model in SPSS.
SPSS Statistics, a software suite utilized for data manipulation, analysis, and data presentation, has a noteworthy history [4]. Its inception dates back to the late 1960s, when a group of Stanford University graduate students first developed it. Originally dubbed as the "Statistical Package for the Social Sciences," the acronym SPSS emerged from its initial focus on social science research. However, as time progressed, the software broadened its scope to encompass the realms of hard sciences and business markets, prompting a rebranding to "Statistical Product and Service Solutions."
In a significant development, IBM acquired SPSS in 2009, leading to a transformation of its name to "IBM SPSS Statistics" [5]. At the time of my last knowledge update, which was in September 2021, the most recent iteration of the SPSS Statistics traditional software license was identified as SPSS Statistics Version 29.

The SPSS Interface
SPSS primarily utilizes two distinct windows: the Data Editor and The Viewer. The Data Editor, as depicted in Figure 1 serves as the platform for data input and the execution of statistical operations.

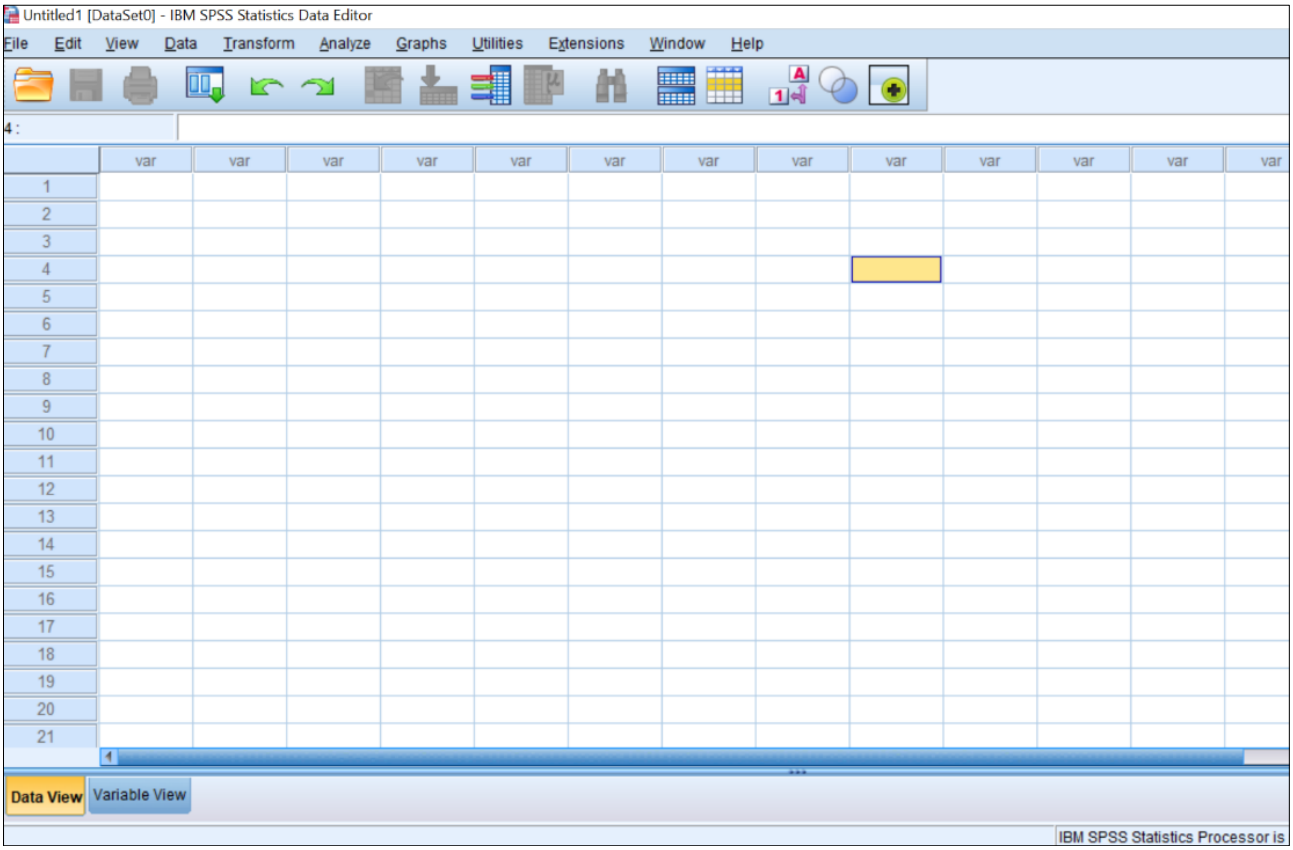


Fig 1: The Data Editor

The Viewer, as illustrated in Figure 2, is the interface where the outcomes of any analysis are displayed. Furthermore, various supplementary windows, such as the SPSS Syntax Editor, can be activated, granting users the ability to input SPSS commands manually [6].

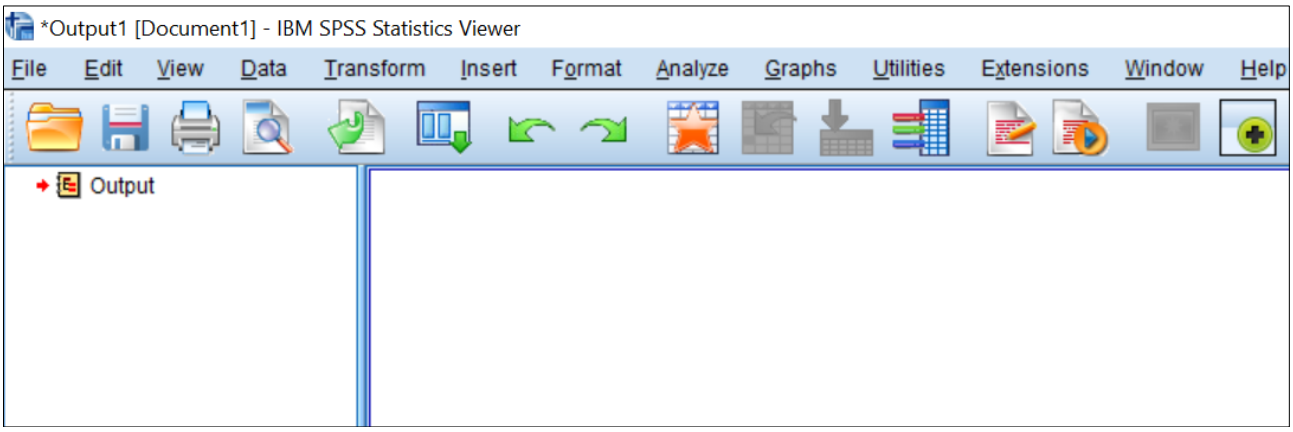
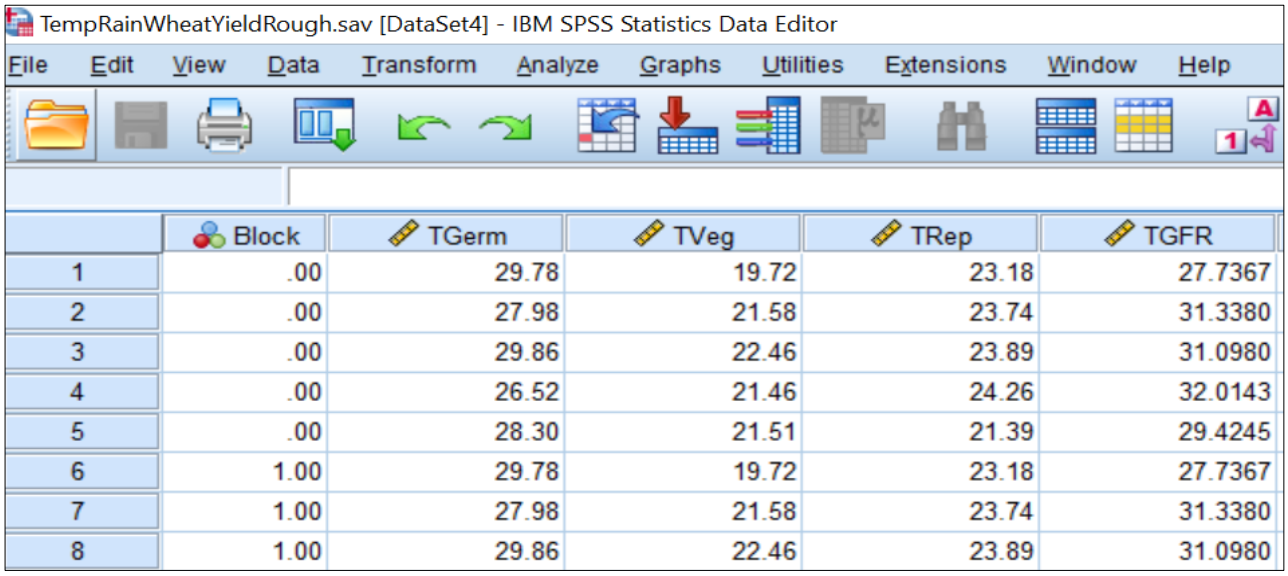


Fig 2: The Viewer

The Data Editor in SPSS offers two distinct perspectives – the Data View and the Variable View. By default, the Data View, where data can be input and reviewed, is readily accessible. Users can effortlessly switch between these views by simply clicking on the tabs located at the bottom left corner of the screen.

In the Data View, the spreadsheet serves as the canvas for entering and visualizing data values. Initially, SPSS opens a blank Data Editor, often labeled as "Untitled1." This arrangement is designed such that each row corresponds to data pertaining to a single entity, while each column represents a distinct variable. Notably, there's no differentiation between independent and dependent variables; both types should be placed in separate columns.

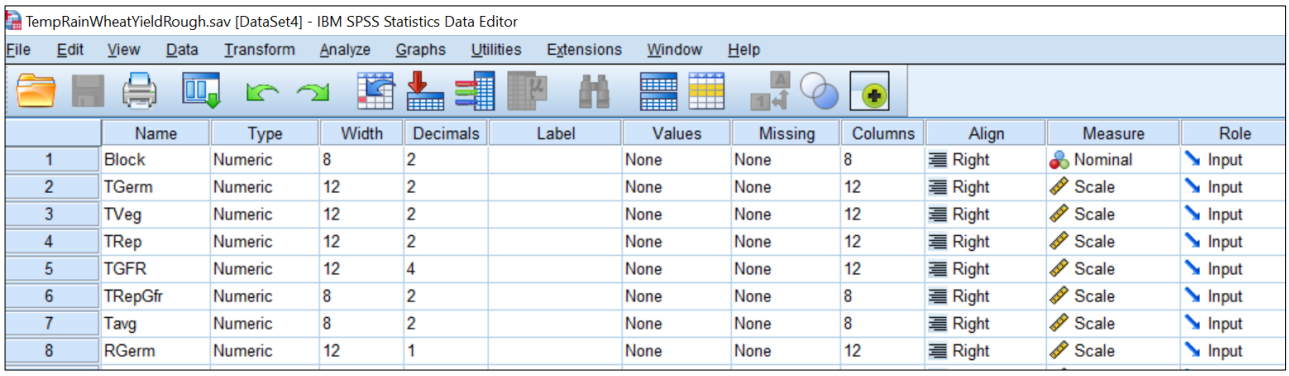
Moreover, numerical data entries are automatically right-aligned within cells, while text (string) entries are left-aligned, enhancing data clarity and interpretation.



	Block	TGerm	TVeg	TRep	TGFR
1	.00	29.78	19.72	23.18	27.7367
2	.00	27.98	21.58	23.74	31.3380
3	.00	29.86	22.46	23.89	31.0980
4	.00	26.52	21.46	24.26	32.0143
5	.00	28.30	21.51	21.39	29.4245
6	1.00	29.78	19.72	23.18	27.7367
7	1.00	27.98	21.58	23.74	31.3380
8	1.00	29.86	22.46	23.89	31.0980

Fig 3: The Data View

In the Variable View, you have the capability to define and observe variable types. The initial stage of data entry involves creating variables through the 'Variable View' in the data editor and subsequently inputting your data using the 'Data View.' In the Variable View, each row corresponds to a variable, and you can establish the attributes of a specific variable by providing information in the designated columns.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Block	Numeric	8	2		None	None	8	Right	Nominal	Input
2	TGerm	Numeric	12	2		None	None	12	Right	Scale	Input
3	TVeg	Numeric	12	2		None	None	12	Right	Scale	Input
4	TRep	Numeric	12	2		None	None	12	Right	Scale	Input
5	TGFR	Numeric	12	4		None	None	12	Right	Scale	Input
6	TRepGfr	Numeric	8	2		None	None	8	Right	Scale	Input
7	Tavg	Numeric	8	2		None	None	8	Right	Scale	Input
8	RGerm	Numeric	12	1		None	None	12	Right	Scale	Input

Fig 4: The Variable View

Data Entry

For the purpose of data manipulation and analysis within SPSS, data is typically input into the software as a data file. This data can be entered directly within SPSS or imported from various external sources. SPSS has the capability to read data stored in IBM SPSS Statistics data files, identifiable by the. sav file extension. Additionally, it offers compatibility with data files from diverse origins, including spreadsheet applications like Microsoft Excel, database applications such as Microsoft Access, and text files.

In the Data View spreadsheet, data values can be conveniently entered. The Variable View spreadsheet is where variables are defined. Each variable's characteristics are specified in a separate row within this spreadsheet. Notably, when data is entered beneath a column in the Data View, the default name of that column automatically populates a row in the Variable View. This streamlined process simplifies data entry and variable definition in SPSS. The features of the variables can be changed by entering information in the different columns in Variable View. The following columns are provided for changing the characteristics of the variables:

1. **Name:** In the Variable View, you can designate the name of each variable within a dedicated column. These names will subsequently appear at the top of their corresponding columns in the Data View, serving as identifiers for the variables in your dataset. Variable names can consist of up to eight alphanumeric characters, but they must commence with a letter. When naming a variable, you may utilize underscores (-), but it's important to note that hyphens (-), ampersands (&), and spaces are not permissible. Furthermore, variable names are not case-sensitive, providing flexibility in how they are represented within SPSS.
2. **Type:** SPSS offers the flexibility to define different types of variables to suit your data. These variable types can include numbers, strings, dates, currencies, and more. When you enter variable values in a column of the Data View, SPSS automatically assigns a default variable type based on the data it detects. However, you can modify this variable type as needed. To change the variable type, you can simply select the corresponding entry in the second column of the Variable View. Then, by clicking on the three-period symbol (. . .) located on the right-hand side of the cell, you can access the Variable Type dialog box. Within this dialog box, you are presented with a range of data types to choose from, including various formats for numerical data, dates, currencies, and more. This feature allows you to precisely tailor the variable type to match the nature of your data in SPSS.

3. **Width:** When a new variable is created, the default type is numeric of width 8, i.e., it can store up to 8 digits. You can modify this value by directly entering a new number in the "Width" column.
4. **Decimals:** This setting specifies the precision for displaying the number of digits to the right of the decimal place. The default setting is 2 and can be changed by a new number in Decimals column.
5. **Label:** There are some restrictions for naming variables such as limit of 8 characters and cannot contain spaces. But during the display of a variable in a chart or in a result of statistical analysis, the full name of a variable is desired. If so, a label may be attached to a variable. These variable labels are valuable for keeping users informed about the significance of variables, and they can be included in the output generated from statistical analyses.
6. **Values:** Values in SPSS are used to associate descriptive labels with different categories of a categorical variable. In the case of categorical variables, each category is assigned an integer code, and the variable is defined as numeric. To set this up, you can click on the corresponding cell in the sixth column of the Variable View, which prompts the appearance of the three-period symbol. Clicking on this symbol opens the Value Labels dialog box, allowing you to assign labels to the category codes. For example, consider a dataset featuring a categorical variable called "Block," representing different Agriculture Development blocks in the Patiala region, with eight distinct values or categories. Clicking on the three-period symbol triggers the opening of the dialog box, as depicted in Figure 5. Within this dialog, numerical codes (e.g., 0 for "Bhunerhedi," 1 for "Ghanour," and so forth) can be defined to represent each block category effectively.

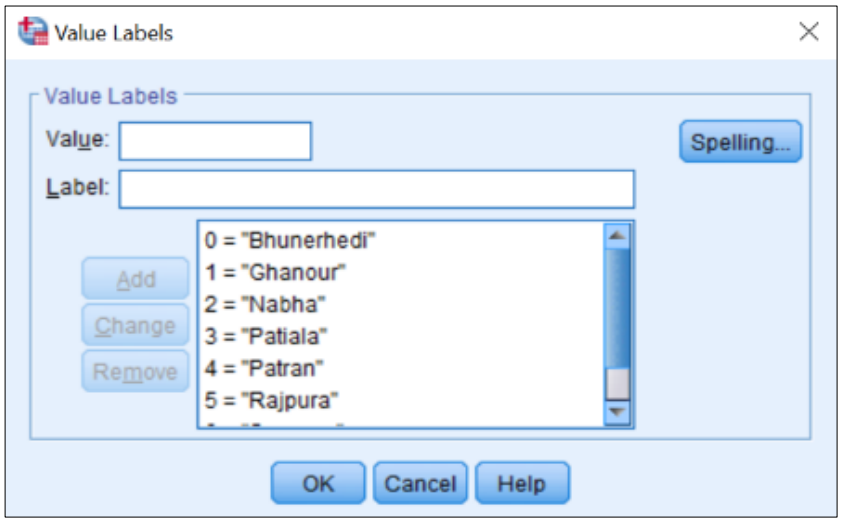


Fig 5: Setting Labels for Categorical Variable – Block

7. **Missing:** In the Data View, a period (.) indicates a missing value to tell SPSS not to treat those data as real, whenever data analyses are conducted. In the Variable View window, under "Missing" that currently there are "none" missing values assigned [7]. IF -111 is to be assigned as missing value for variable – Yield, then select the "None" cell under "Missing" in the "Yield" row, and click the ". . . " button. This will open a "Missing Values" box. Select "Discrete missing values" and enter -111, as below. Then click "OK".

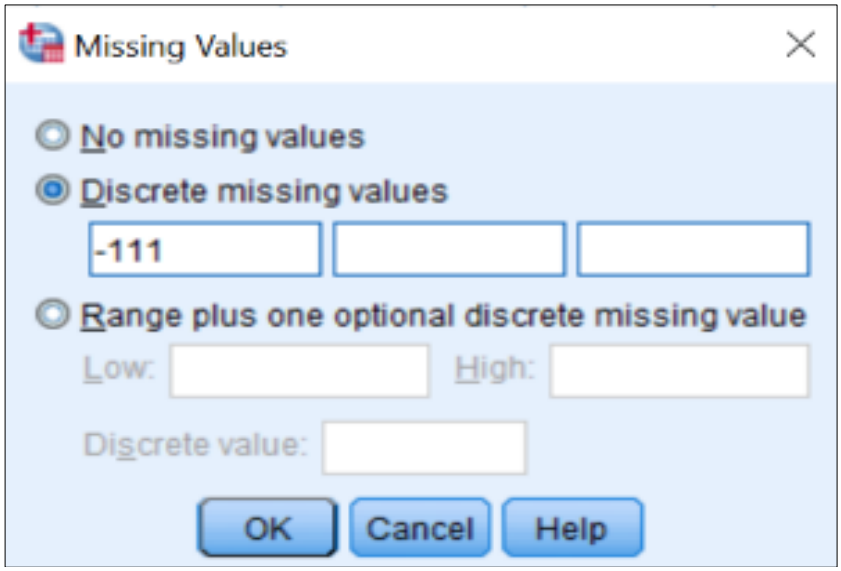


Fig 6: Figure Caption

8. **Columns:** This parameter defines the width of the column associated with the variable in the Data View. By default, numerical variables have a cell width of eight. When the specified "Width" value exceeds the "Columns" value, it's possible that only a portion of the data entry may be visible in the Data View. You can adjust the cell width using the same method employed for altering data entry widths, or you can simply resize it by dragging the column boundary accordingly.
9. **Align:** This column allows you to modify the default alignment of the data within the corresponding column.
10. **Measure:** This specifies the measurement scale of a variable. For most of the types three measures – Nominal, Ordinal and Scale are specified. A variable in SPSS can fall into two broad categories: categorical or continuous, and it may possess various levels of measuremen [8].

a) **Categorical:** Data characterized by a finite number of distinct values or categories is referred to as categorical data. Categorical variables in SPSS can take the form of string data (alphanumeric) or numeric variables that utilize numeric codes to represent these categories.

i) **Nominal:** Categorical data without a natural or inherent order among the categories is typically referred to as nominal data. For example, the yield data is collected from different Agriculture Development Blocks and the category of one block is not higher or lower than the other.

ii) **Ordinal:** Categorical data that exhibits a meaningful order among its categories, yet lacks a quantifiable or measurable distance between those categories, is often termed as ordinal data. For instance, in the case of a variable like "Yield Class," which categorizes yield as High, Medium, or Low, there exists an order among the values (High > Medium > Low), but you cannot calculate a precise numerical distance between the

b) **Scale:** Data that is measured on either an interval or ratio scale is characterized by values that not only convey the order of those values but also quantify the precise distance between them. For instance, in the case of salary data, where a salary of 82,500 is indeed higher than a salary of 62,150, and the exact distance between these two values is 20,350, this data is considered to be on an interval or ratio scale.

Statistical Models

The overarching objective of data analysis is to gain valuable insights from the provided data. To facilitate this process, statistical models serve as a valuable framework. These models are instrumental in identifying connections between variables and comprehending the impact of these variables on a broader system, whether they are operating individually or in tandem. Moreover, statistical models aid in making predictions and evaluating the degree of uncertainty associated with these predictions. Typically, these models are represented as a set of equations or mathematical formulas, which articulate how certain aspects or all aspects of the data could have been generated [9].

Stepwise Multiple Linear Regression

Regression analysis involves using existing data to predict other values[10]. This process entails creating a model that maps past values of the output variable to input variables. Once this model is established, it is employed to forecast the output variable's value based on the input variables in a manner that minimizes prediction errors [2]. Regression analysis encompasses several variations, including Simple Linear Regression, Multiple Linear Regression, Non-linear Regression, and Logistic Regression. Simple Linear Regression involves one dependent variable and one independent variable. The basic relationship between two variables, X and Y, can be expressed as:

$$Y = a + bX \quad (1)$$

The values of a and b are chosen so that the error is minimum. To calculate the error, squared difference of the predicted and the actual value is found.

In Multiple Linear Regression, there is one dependent variable and multiple independent variables. If there is a dependent variable 'Y' and three independent variables 'X1,' 'X2,' and 'X3,' the equation takes the form:

$$Y = a + b_1(X_1) + b_2(X_2) + b_3(X_3) \quad (2)$$

Simple Linear Regression describes a line in two-dimensional space, whereas Multiple Regression extends to (n+1)-dimensional space, where 'n' represents the number of independent variables. When constructing complex models with several predictors, the selection of appropriate predictors is crucial. SPSS provides various methods for predictor selection to aid in this process.

Hierarchical Regression is a method that involves entering predictors in a specific order determined by the researcher based on prior knowledge and experience. Another approach is the Enter method (Forced Entry), which involves adding all predictors simultaneously to the model. When the order of predictor inclusion is solely determined by mathematical criteria, it's referred to as Stepwise Regression.

Stepwise Multiple Linear Regression can be of two types: Forward or Backward. In the Forward method, an initial model contains only a constant 'a.' Then, the next predictor that best predicts the dependent variable is identified. The predictor with the highest simple correlation with the dependent variable is chosen. If this predictor significantly enhances the predictive capacity of the model, it is retained, and the search for the next predictor continues. The second predictor is selected based on having the largest semi-partial correlation with the dependent variable. For instance, if the first predictor explains 25% of the variation in the dependent variable, the remaining 75% unexplained variation is considered. The next predictor selected is the one that can explain the largest part of this remaining 75%. Therefore, semi-partial correlation measures how much 'new variance' in the dependent variable can be explained by each remaining predictor. This process continues, adding predictors that contribute significantly to the model's predictive power.

The Backward method follows the opposite approach. Initially, all predictors are included in the model, and then, based on the significance value of the t-test for each predictor, they are removed one by one. The model is re-estimated with the remaining predictors, and the contribution of these remaining predictors is reassessed until no further predictors meet the removal criterion [6].

To perform Regression Analysis in SPSS, follow these steps

- Open your dataset in SPSS.
- Navigate to the "Analyze" menu.
- From the "Analyze" menu, select "Regression" and then choose "Linear."

This action will open the Linear Regression dialog box as shown in Figure 7, which allows you to configure your analysis. Inside the dialog box:

- On the left-hand side, you'll see a list of all the variables in your dataset.
- Locate the space labeled "Dependent" and place your target or outcome variable in this field.
- Similarly, identify the space labeled "Independent" and insert your predictor variables in this field.

Additionally, you can utilize the "block controls" within the Linear Regression dialog box to build a series of regression models, specifying which variables should be included at each stage or block of the analysis according to your research design or criteria. This approach provides flexibility and control in setting up and executing your linear regression analysis in SPSS.

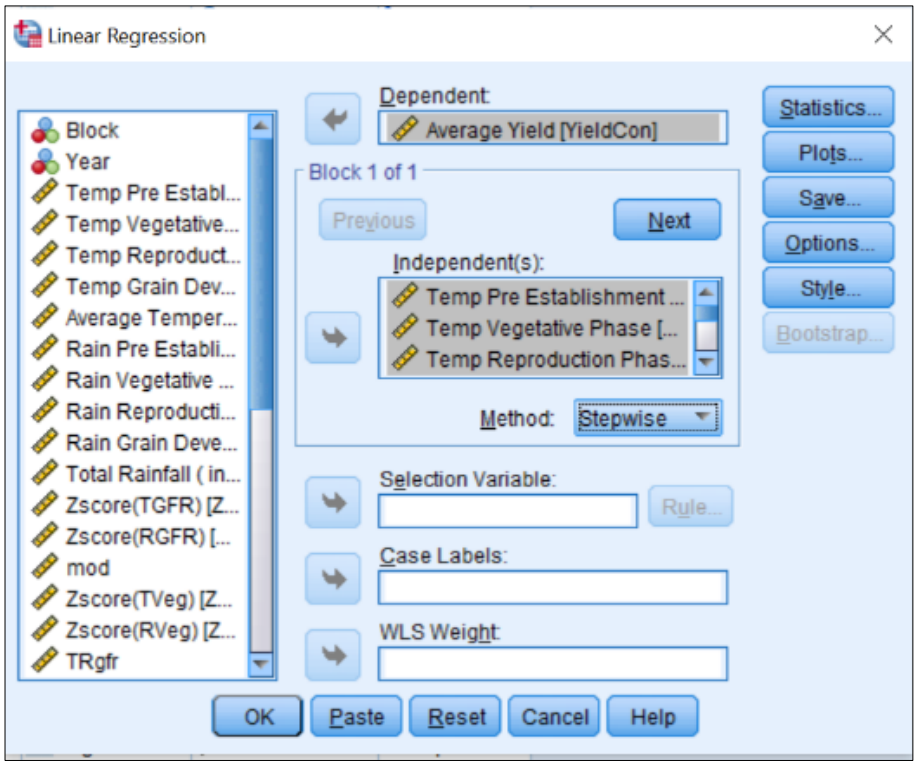


Fig 7: Multiple Linear Regression in SPSS

By default, the method selected is "Enter," indicating that all independent variables within the block will be included simultaneously in the regression equation. The other options are – Stepwise, Remove, Forward and Backward.

Hierarchical Regression can also be used. In this case, specify the target variable in the same way as described above. Then choose the independent variable that is to be included in the first phase. After specifying the first block in the hierarchy, you proceed to the next. To signal the software that a new set of predictors should be specified, you can click on the "Next" button. This action clears the "Independent(s)" box, allowing for the entry of new predictors. The top of the box indicates, for example, "Block 2 of 2," indicating that you are currently working in the second block of the two you have specified so far. Likewise, you can add the next predictor in the hierarchy by clicking "Next," and you can navigate between blocks using the "Previous" and "Next" buttons.

The "Selection Variable" option permits cross-validation of regression outcomes. This means that values meeting the specified rule for a selection variable are utilized in the regression analysis, while the resulting prediction equation is applied to other cases. Consequently, you can evaluate regression on cases not used in the analysis or apply the equation derived from one subgroup of data to other groups.

While SPSS Statistics provides standard regression output by default, additional statistics can be requested through the Statistics dialog box. The Plots dialog box is used to generate various diagnostic plots, including residual plots, often used in regression analysis. The Save dialog box enables the addition of new variables to the data file, containing statistics such as predicted values, residuals, and influence measures. Lastly, the Options dialog box provides control over criteria during stepwise regression and choices for handling missing data. By default, SPSS Statistics excludes a case from regression if one or more values are missing for the variables used in the analysis. The result of the Regression Analysis can be interpreted by statistics explained in Section 6.

Statistics Used

Null Hypothesis

Hypothesis test is used when an inference has to be made about a population from a sample. The purpose is to test whether random chance might be responsible for an observed effect [11]. Hypotheses in statistical analysis typically fall into two categories: the null hypothesis (H0) and the alternative hypothesis (H1). A null hypothesis always states that no effect is present and if any effect is observed, it is due to random chance. Conversely, the alternative hypothesis is derived from a specific theory or prediction and typically asserts the presence of an effect or relationship.

After setting up the hypothesis, the value of test statistics is calculated using the sample observations. The test statistic is calculated as the ratio of variance explained by an effect to the variance not explained by an effect. The value of test statistic is used to reject or accept the null hypothesis.

$$Test\ stastic = \frac{Variance\ explained\ by\ an\ effect}{Variance\ not\ explained\ by\ an\ effect} \quad (3)$$

Level of Significance

Level of significance is the probability to reject H_0 when H_0 is actually true. Usually, the null hypothesis is not just simply rejected or accepted; rather it is rejected when the sampling result (observed evidence) has less than a certain probability of occurring, if H_0 is true. This probability is called level of significance. This means that H_0 is rejected with a certain significance level or confidence level; that level may be 1%, 5% or 10%.

$$Confidence\ Level = 100 - Level\ of\ Significance \quad (4)$$

So, if the level of significance is 1%, 5% or 10%, the level of confidence is 99%, 95% or 90%, respectively. Usually, 5% significance is chosen which implies 95% confidence is selected.

The p-value is not the probability of the null hypothesis being true; instead, it quantifies the strength of evidence against the null hypothesis. It represents the probability of observing a test statistic as extreme as, or more extreme than, the one obtained from your data, assuming that the null hypothesis is true. The decision to reject the null hypothesis is based on whether the p-value is less than or equal to a pre-defined significance level (α), such as 0.05 (5%). The critical value, specific to the chosen significance level and statistical test, is used to determine whether the calculated test statistic falls into the 'critical region' (leading to the rejection of the null hypothesis) or the 'non-critical region' (indicating a failure to reject the null hypothesis).

Covariance and Correlation

When analyzing the relationship between two variables, it's essential to assess whether changes in one variable correspond to similar changes in the other variable. In this context, when one variable deviates from its average (mean), we anticipate that the other variable will also deviate from its mean in a comparable manner. To quantify this relationship, we use the concept of covariance between two variables, denoted as σ_{xy}, and it is calculated using the following formula:

$$Covariance = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \quad (5)$$

Here, $\bar{x}$ represents the mean of the observations for variable x, and $\bar{y}$ represents the mean of the observations for variable y. Covariance can assume values ranging from negative infinity to positive infinity. A positive covariance indicates that an increase in the value of one variable corresponds to an increase in the value of the other variable. Conversely, negative covariance suggests that an increase in one variable corresponds to a decrease in the other variable. A covariance of 0 signifies no discernible relationship between the variables.

However, one limitation of covariance as a measure of the relationship between variables is its dependence on the scales of measurement used. Therefore, covariance is not a standardized measure. To address this limitation, Karl Pearson introduced the coefficient of correlation, denoted as R, which standardizes the covariance:

$$R = \frac{Covariance(x,y)}{Stddev(x).Stddev(y)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)\sigma_X.\sigma_Y} \quad (6)$$

Here, σ_X represents the standard deviation of variable x, and σ_Y is standard deviation of variable y. represents the standard deviation of variable y. Karl Pearson's coefficient of correlation is not influenced by the scales of measurement and yields values within the range of -1 to +1.

- A coefficient of +1 indicates a perfect positive correlation, implying that as one variable increases, the other increases proportionately.
- Conversely, a coefficient of -1 indicates a perfect negative correlation, signifying that as one variable increases, the other decreases proportionately.
- A coefficient of 0 implies no linear relationship whatsoever, suggesting that when one variable changes, the other remains unaffected.

Coefficient of Determination (R squared)

The Coefficient of Determination represents the fraction of variability in one variable that can be explained by the other variable. It is determined by squaring the value of the correlation coefficient, denoted as "r." For example, if r = 0.5, then r^2 = 0.25, signifying that 25% of the variation in one variable can be predicted or attributed to the knowledge of the values of the other variable. In the context of regression analysis, the value of r^2 signifies the proportion of the variance in the dependent variable that can be clarified by the independent variables.

Sum of Squares

In regression analysis, we define three types of sum of squares (ss): Total Sum of Squares, Regression Sum of Squares, and Residual Sum of Squares.

1. **Total Sum of Squares (Total SS):** This is calculated as the sum of the squared deviations of the observed values of the variable from the sample mean. The formula for Total SS is as follows:

$$Total\ ss = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \quad (7)$$

Here, $\bar{x}$ represents the sample mean.

2. **Regression Sum of Squares (Regression SS):** This is computed as the sum of the squared deviations of the predicted values of the variable from the sample mean. The formula for Regression SS is:

$$Regression\ ss = \sum_{i=1}^{n}(\hat{x}_i - \bar{x})(\hat{x}_i - \bar{x}) \quad (8)$$

Here, $\hat{x}_i$ represents the predicted values based on the regression line, and $\bar{x}$ is the sample mean.

3. **Residual Sum of Squares (Residual SS):** Residual represents the differences between each observed data point and the value predicted by the regression line. Residual ss is defined as the sum of the squared deviations of the predicted values of the variable from the observed values.

$$Residual\ ss = \sum_{i=1}^{n}(x_i - \hat{x}_i)(x_i - \hat{x}_i) \quad (9)$$

Here, x_i represents the observed values, and $\hat{x}_i$ represents the predicted values based on the regression line.

An important relationship is that Total SS is equal to the sum of Regression SS and Residual SS. This relationship is fundamental in regression analysis and is expressed as:

Total SS = Regression SS + Residual SS

Understanding these components helps in assessing the goodness of fit of a regression model and understanding how much of the variability in the dependent variable can be explained by the independent variable(s) represented by the regression model.

Degree of Freedom

Degree of Freedom (df) is a parameter that allows test statistic distributions to adjust to different sample sizes and number of groups. It is the number of values in the final calculation of a statistic that are free to vary. df is calculated differently for different test statistic. In Regression, df is calculated for the Model, the Residual and the Total as shown in the following equations:

$$df\ (Regression) = Number\ of\ Regressors - 1 \quad (10)$$

$$df\ (Residual) = Sample\ size - Number\ of\ Regressors \quad (11)$$

$$df\ (Total) = df\ (Regression) + df\ (Residual) \quad (12)$$

Mean Square

Mean Square (MS) is defined as the Sum of Squares divided by the Degree of Freedom.

$$Mean\ Square = \frac{Sum\ of\ Squares}{Degree\ of\ Freedom} \quad (13)$$

F-score

F-score is calculated by dividing the Mean Square of Regression with the Mean Square of Residue.

$$F - score = \frac{MS\ (Regression)}{MS\ (Residual)} \quad (14)$$

Conclusion

In conclusion, this paper has outlined the significant capabilities of SPSS Statistics in the realm of machine learning and regression modeling, particularly when handling extensive datasets. It has provided insights into how SPSS enables users to apply diverse techniques, such as Hierarchical and Stepwise Regression, to extract valuable insights from substantial data collections. The examination of SPSS's core interface components, namely the Data Editor and the Viewer, has highlighted their essential functions. Additionally, the paper has discussed data entry processes, the comprehensive options available in Variable View for defining variable characteristics, and the importance of fundamental statistical concepts like the null hypothesis and level of significance. In essence, SPSS emerges as an indispensable tool for data mining, allowing the creation of robust regression models for predictive analyses while rigorously evaluating model accuracy using metrics such as sum of squares, mean absolute error, root mean squared error, and more.

References

1. Attwal KPS. Design and Development of a Model for Wheat Yield Prediction Using Data Mining. Patiala: Punjabi University; c2020.
2. Attwal KPS, Dhiman AS. Investigation and Comparative Analysis of Data Mining Techniques for the Prediction of Crop Yield. International Journal of Sustainable Agricultural Management and Informatics. 2020;6(1):43-74.
3. Attwal KPS, Dhiman AS. Exploring Data Mining Tool – WEKA and Using WEKA to Build and Evaluate Predictive Models. Advances and Applications in Mathematical Sciences. 2020;19(6):451-469.
4. Landau S, Everitt BS. A Handbook of Statistical Analysis Using SPSS. CRC Press; c2004.
5. George PM. IBM SPSS Statistics 23 Step by Step: A Simple Guide and Reference. Routledge; c2016.
6. Field A. Discovering Statistics Using IBM SPSS. Sage; c2013.
7. Hayes AF. Introduction to Mediation, Moderation and Conditional Process Analysis. The Guilford Press; c2018.
8. IBM. IBM SPSS Statistics 25 Brief Guide. IBM; c2019.
9. National Research Council. Frontiers in Massive Data Analysis. Washington: The National Academic Press; c2013.
10. Edelstein HA. Introduction to Data Mining and Knowledge Discovery. Potomac, MD, USA: Two Crows Corporation; c2005.
11. Bruce P, Bruce A. Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly; c2017.
12. Hayes AF, Rockwood NJ. Regression-Based Statistical Mediation and Moderation Analysis in Clinical Research: Observations, Recommendations, and Implementation. Behaviour Research and Therapy; c2016.
13. Rahman S, Xu H. A Univariate Dimension-Reduction Method for Multi-dimensional Integration in Stochastic Mechanics. Probabilistic Engineering Mechanics. 2004;19(4):393-408.
14. Ribeiro MHDM, Coelho LdS. Ensemble Approach Based on Bagging, Boosting and Stacking for Short-term Prediction in Agribusiness Time Series. Applied Soft Computing. 2020;86:105837.
15. Sagi O, Rokach L. Ensemble Learning: A Survey. Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018;8(4):e1249.
16. Suthaharan S. Support Vector Machine. In: Springer, editor. 2016;36:207-235.
17. Wu XK, Zhu X, Gong Q, Zhang J, Yu P, Xindong W, *et al*. Top 10 Algorithms in Data Mining. Knowledge and Information Systems. 2008;14:1-37.
18. Yang P, Liu W, Zhou BB, Chawla S, Zomaya AY. Ensemble-based Wrapper Methods for Feature Selection and Class Imbalance Learning. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Heidelberg; c2013.
19. Zhang GP. Neural Networks for Data Mining. In: Data Mining and Knowledge Discovery Handbook. New York, USA: Springer; 2010. p. 419-444.
20. Hyma J, Varma PS, Kumar SVSN, Salini GR. Heterogeneous Data Distortion for Privacy-Preserving SVM Classification. In: Smart Intelligent Computing and Applications. Singapore: Springer; 2019. p. 459-468.
21. Osorio DL, Lombard G. Descriptive Data Mining. Singapore: Springer; c2019.