THOMPSON RIVERS UNIVERSITY

Evolutionary algorithm applied to constrain coefficients of dummy variables

By

Sarah Chopra

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science in Data Science

KAMLOOPS, BRITISH COLUMBIA

April, 2023

SUPERVISOR

Dr. Mateen Shaikh

ABSTRACT

Simulated datasets with categorical and continuous variables are analyzed to understand how the coefficients of dummy variables in a regression model respond to different sample sizes, change in a coefficient of one level of one categorical variable, and change in variance of residuals. An evolutionary algorithm is employed to reduce the model's complexity by replacing coefficients of different levels of categorical variables with coefficients of other levels of the same categorical variable and simplifying the model. The algorithm is tested on datasets with five levels for two categorical variables to determine which models are selected. The approach is also applied to American income data. The algorithm finds that several levels of the categorical variables have the same effect on the regression; hence, those levels are similar in the model.

Key Words: Evolutionary algorithm; dummy variables; AIC; American income data; linear model].

Contents

1 Introduction **1**

1.1 Supervised machine learning algorithm 1

1.2 Linear Regression 2

1.2.1 Categorical Variables 4

2 Literature Review **5**

2.0.1 Simple linear regression model 5

2.0.2 Evolutionary Algorithm 8

2.0.3 Model Dimension Reduction 9

2.0.3.1 Feature Selection 9

2.0.3.2 Lasso regression 10

2.0.3.3 Principal component analysis 11

3 Methodology 12

3.0.1 Simulated Data 12

3.0.1.1 Varying true coefficient of one level of one dummy variable 13

3.0.1.2 Varying the sample size 14

3.0.1.3 Varying variance of residual 14

3.0.2 Collinearity and least square method 14

3.0.3 Use of Linear regression 16

3.0.4 Evolutionary Algorithm 17

3.0.4.1 Census data analysis 20

4 Results 23

4.0.1 Varying X variable's Coefficient for E 23

4.0.2 Parameters Vs sample size 30

4.0.3 Parameters Vs variance of residuals 35

4.0.4 Census Data results 43

5 Discussion 44

5.0.1 Selecting the most important categorical variable first . 44

5.0.2 Comparing Lasso to Evolutionary Algorithm 45

5.0.3 Algorithm Limitation and big data application 45

5.0.4 Use of CV to find the threshold 46

List of Figures

2.1 a) Represents fitted vs residual plot, b) QQ plot (left-to-right) 7

3.1 An evolutionary algorithm 20

3.2 Coefficient (Intercept and X1) Vs K 21

3.3 Coefficient (Intercept and X1) Vs K 21

3.4 Coefficient (Intercept and Z) Vs K 21

3.5 Coefficient (Intercept and Z) Vs K 22

4.1 Coefficient (Intercept and Z) Vs K 23

4.2 Coefficient of X variable's 5th level(E) Vs K 26

4.3 Coefficient of X variable's 2nd,3rd and 4th (B,C,D) Vs K . . 27

4.4 Coefficient of Y variable's 2nd,3rd, 4th and 5th (G,H,I,J) Vs K 30

4.5 Coefficient (Intercept and Z) Vs sample size 31

4.6 Coefficient of Y variable's 2nd,3rd, 4th and 5th (G,H,I,J) Vs sample size . . . 33

4.7 Coefficient of Y variable's 2nd,3rd, 4th and 5th Vs sample size 35

4.8 Coefficient (Intercept and Z) Vs variance of residuals 36

4.9 Coefficient of X variable's 2nd,3rd and 4th Vs variance of residuals . . . 39

4.10 Coefficient of Y variable's 2nd,3rd, 4th and 5th Vs variance of residuals . . . 41

List of Tables

3.1 True Coefficients . . . 13

3.2 Sample simulated data(with collinearity) . . . 15

3.3 Sample simulated data (without collinearity) . . . 16

3.4 Example . . . 18

4.1 Intercept comparison (case I) . . . 24

4.2 Z comparison (case I) . . . 25

4.3 X2 comparison (case I) . . . 25

4.4 X comparison (case I) . . . 27

4.5 Y comparison (case I) . . . 29

4.6 Intercept comparison (case II) . . . 31

4.7 Z comparison (case II) . . . 31

4.8 X2 comparison (case II) . . . 32

4.9 Y comparison (case II) 34

4.10 Intercept comparison (case III) 36

4.11 Z comparison (case III) 36

4.12 X comparison (case III) 38

4.13 Y comparison (case III) 40

Chapter 1

Introduction

Most earlier research has indicated that humans do not combine information as good as regression models (1). Humans tend not to combine all the information from the independent variables, which results in significant information loss while predicting. It is crucial to analyze all the independent variables in the dataset before predicting an unknown variable (1). This project will strive to lower the model's dimensionality while attempting to capture most of the information in the independent variables.

1.1 Supervised machine learning algorithm

A supervised machine learning algorithm is an algorithm that works on labeled datasets. A labeled dataset is a dataset that contains correct outputs (2). A supervised algorithm trains using a training dataset and generates a model. The model is then used on the test dataset to make predictions.

The concern with supervised machine algorithms is that the model might only succeed on test data if all the scenarios in test data are encountered while training the model (2).

1.2 Linear Regression

Linear regression is a supervised machine learning algorithm. It uses training data to train the model. As a supervised machine learning algorithm, it uses labeled datasets with the correct values for the predicted variable. The predicted variable is also called the target variable, which is to be predicted. They are also called dependent variables or regressand (3). The other type of variable is the independent variable(s). These variables do not depend on any other variable in the dataset and are used to make predictions about predictor variables (3). Linear regression is a popular algorithm because it interprets the relationship between variables simply and easily.

In linear regression, the independent variables are defined using X.

X = X_1,X_2...X_n, and f(X) as the predictor variable takes the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

The equation above is the linear model equation (3).In the equation β_0 is the intercept. This constant value of the intercept denotes the average value of the target variable, keeping all the independent variables zero. The coefficient of X_j, β_j where j = 1 represents the coefficient of the first independent variable. It denotes the average change in the target variable for a unit increase in the first independent variable. A method known as least squares estimates the coefficients in a linear model (4). The least square method defines a line of best fit that tries to accommodate all the points in the dataset. The least square method minimizes the residual sum of squares, and it is formulated as follows:

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Including all independent variables in the model might seem helpful in achieving the best predictions. Including many independent variables in the model leads to overfitting, where the model becomes too complex. Also, if one of the independent variables is random and does not contribute to the betterment of the model; in that case, this variable may bring down the prediction power of the model. Therefore, comparing different models and trying to remove or add independent variables is essential to make better predictions. The Akaike Information Criterion (AIC) values can be used to select between different possible linear regression models (5). AIC is calculated as :

$$AIC = Nlog(SS_{error}/N) + 2K$$

Here, N is the sample size, SSerror is the sum of squares of errors for the model, and K is the number of independent variables fitted in the model. It can be seen that with the increase in K, AIC increases, which means if there is an increase in the number of parameters that fit into the model, the value for AIC increases. If we keep the parameters fixed and increase the sample size alone, the AIC value is expected to decrease due to log(1/N). With increased sample size, log(1/N) would become a larger negative value. This aspect of the datasets will also be evaluated. It should be noted that a lower AIC value indicates a better model, and conversely, a higher AIC value indicates a worse model (5).

1.2.1 Categorical Variables

A categorical variable in data means a variable that can only have one value out of a fixed number of possible values (6). For example, the inbuilt iris dataset in R has a specific feature called species. There are three kinds of species, namely virginica, setosa, and versicolor. All the data points in the dataset have one of these three values in the species column. While working with this dataset, if species is used as an independent variable, in order to apply linear regression, dummy coding needs to be applied. Under dummy coding, two indicator variables are created for whether the iris is virginica, setosa, or versicolor. If a record is of type setosa, two indicator variables will have values 0,1, respectively. If a record is versicolor, the indicator values will be 0 and 0, respectively. With a dummy variable, the regression for this categorical variable takes the form:

$$Y = b_0 + b_1 setosa + b_2 virginica$$

In the Iris flower example, the three species, virginica, setosa, and versicolor, are represented in the equation. The reference category is versicolor and corresponds to the case that all indicators (setosa and virginica) are 0. All other coefficients are calculated as changes from this reference variable. Versicolor is taken as reference and is effectively represented in the model by the intercept, b_0. For this study, the R compiler chooses the reference value.

Chapter 2

Literature Review

2.0.1 Simple linear regression model

The linear regression model mainly focuses on solving two problems. The first problem is of defining the relationship between variables. The relationship between any two variables can also be measured by a quantity called correlation (3). In linear regression, a straight line can express a linear relationship between two variables. The dataset's points might not be on the line, but their distance from the line is minimized. In linear regression, the best-fit line created after applying the least squares method tries to accommodate all the data points by minimizing the error or residuals, which is the distance between the actual data point and the estimated or predicted point on the line. The slope of the best-fit line defines the relationship between the target variable and the independent variables. A model can identify trends or predict by analyzing the variables' slope or relationship. The equation for the linear model :

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

In the above equation, β_0 denotes the intercept. It gives the mean value of the target variable, given that all independent variables are 0.

A correlation coefficient measures correlation. It can range from -1 to 1 (7). If the correlation coefficient is -1, it indicates a perfect negative correlation meaning both the variables' fluctuation concerning each other is similar but is in opposite directions. Nevertheless, the fluctuation concerning each other is similar and in identical directions for a positive relation. For example, the number of goals by a player in basketball can be positively correlated with practice and height but might be negatively correlated with how lazy the player is.

The second problem linear regression tries to accommodate is the ability to predict. Once a model is created using the training data, the same model is used to predict the test data. The prediction works best if all or most possible scenarios are encountered in the training data.

There are four assumptions of multiple linear regression (8). The first assumption of the four assumptions is the normal distribution of the residuals. If the residuals are not normally distributed, it will distort the regression results (8). The second most crucial assumption is that the target and independent variables have a linear relationship. Non-linear relationships are hard to express in linear regression. For example, the temperature rises as we go toward the equator, given that all other factors are nullified. It means that the shorter the distance to the equator, the more the temperature. Hence, these two variables might have a linear relationship. Understanding whether data is linearly distributed can be crucial in statistical analysis, as it can help inform the appropriate choice of statistical methods and models for analyzing the data.

The third assumption is homoscedasticity (8). It refers to a situation where the variance of residuals is constant for all the data points. A violation of the homoscedasticity assumption can lead to prediction errors, resulting in incorrect inferences and predictions.

The fourth assumption of linear regression is the independence of errors. When the errors are independent, errors and target variables are not correlated. Any individual error term is also not related to any other error term. This assumption of error independence is crucial in various statistical models, as it makes valid inferences about the relationships between variables.

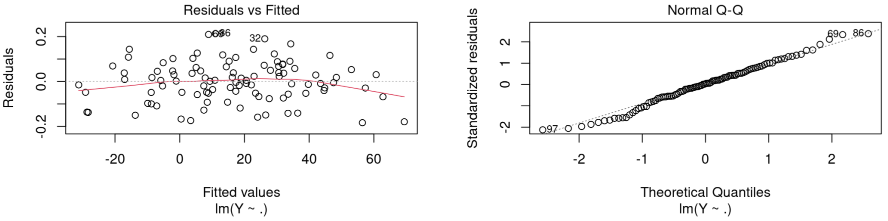

Figure 2.1: a) Represents fitted vs residual plot, b) QQ plot (left-to-right)

Figure 2.1 a) The plot describes the relationship between residuals and fitted values. This plot helps find any outliers in the data and the relationship between residuals and fitted values. It can be seen that there is no relationship between the two quantities. Hence, residuals are independent of fitted values.

Figure 2.1 b) A QQ plot helps determine if the collection of points follows the normal distribution. The comparison is made between expected behavior, seen as the perfect diagonal line on the plot starting from the left down corner to the top right corner, and the actual data represented by black points. The plot shows that data almost follows the normal distribution.

2.0.2 Evolutionary Algorithm

An evolutionary algorithm is a coded set of rules that can optimize a problem. This type of algorithm is very similar to natural selection and the process of evolution (9). The evolutionary algorithm works on population and employs selection, recombination, and mutation. The evolutionary algorithm retains the best features; it creates new solutions by modifying the best features. It is applied in a range of fields ranging from engineering to biology. Every evolutionary algorithm has a fitness or objective function (10). The fitness function is the objective function optimized during the evolutionary process. The fitness function determines how well a particular set of parameters or candidate solution performs on a given problem. In the context of evolutionary algorithms and other optimization techniques, a population is a collection of potential solutions or candidate solutions to a problem. Each member of the population is called an individual, representing a possible solution to the problem being solved. Similarly, the selection is a process in evolutionary algorithms where the fittest individuals from the current population are chosen as references for the next generation. It tries to preserve the best solution during the algorithm's run. All the steps are applied; an evolutionary algorithm converges to an almost optimal solution.

Below steps are general steps in any evolutionary algorithm:

1) Initialize a population of randomly selected solutions.

2) Evaluate how good each of the solutions is.

3) Generate offspring from the best solutions and replace the worst solutions with offspring until a stopping criterion is met.

4) Terminate the code/program if the stop condition is met.

5) Return the best solution from the final population.

2.0.3 Model Dimension Reduction

The dimensionality of a dataset is defined as the number of input features (11). These are the independent variables in a model which help predict the dependent variable. With massive datasets it becomes challenging to manage, store, manipulate, and analyze massive datasets. Model dimension refers to the number of parameters in the statistical model. The complexity or dimensionality of the model can have a significant impact on its performance.

2.0.3.1 Feature Selection

With increased data volume, data quality might decrease. Higher dimension data brings noise and redundancy to the produced models. Hence, it is essential to find a way to reduce the number of independent features. The main idea of feature selection is to evaluate the relationship between each independent feature and the target feature. The independent features with the most vital relationship with the target feature get selected. There are two different types of feature selection which are supervised and unsupervised.

The target variable is known in supervised feature selection, whereas the target variable is unknown in unsupervised feature selection. One type of supervised feature selection is feature-based feature selection. This method uses statistical measures to measure the correlation between independent and output features to determine the best features (independent). While filtering under this type, the relationship between independent variables is not considered; only the relationship between each independent variable and target variable is evaluated to select the best independent features. This method results in model dimensionality reduction.

2.0.3.2 Lasso regression

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression model which uses the L1 technique. The L1 technique adds a penalty value to the model equation. This penalty is equal to the absolute value of the coefficient. Hence, some coefficients become zero post application, and the model becomes sparse (12).

Lasso regression can be expressed by:

$$SSerror + \lambda * (Sum_of_absolute_value_of_coefficients)$$

One of the significant differences between lasso and linear regression is that the linear regression model interprets each category level. However, lasso regression works only on the categorical variable and not separately on categorical variable's levels.

2.0.3.3 Principal component analysis

Principal Component Analysis, or PCA, is a dimensionality reduction technique used to reduce the dimension of large datasets. PCA tries to retain the maximum variation, so there is little effect on prediction. PCA can be described using five significant steps (13). The original dataset's continuous variables are standardized. This is done so that each variable contributes equally to the analysis. Also, PCA is sensitive to the variance of individual features. The reason is that the more the variance of the column or feature, the more those features will dominate over other features and can result in incorrect results. The next major step for PCA is to generate the covariance matrix. The covariance matrix helps us understand how each variable varies concerning others. The generation of principal components takes place post-generating covariance matrix by computing the eigenvectors and eigenvalues. Post-computing eigenvectors,eigenvectors are ordered by eigenvalues to generate principal components in order of significance. Finally, the data is reoriented from the original axes to ones represented by principal components.

Chapter 3

Methodology

3.0.1 Simulated Data

The data used in this project is simulated. The data contains a few different scenarios to understand and interpret the results in an easy and less complex way. The simulated data is concise but expresses the concern of dimensionality well. Evolutionary algorithm is executed over the simulated data. The dataset comprises four distinct columns. Column Z is a continuous variable. The remaining two columns, X and Y, are categorical variables with five levels each. An equivalent number of records are allocated to each level of both categorical variables. The five possible values for X categorical variable are level11(A), level12(B),level13(C),level14(D),level15(E). The five possible values for Y categorical variable are level21(F), level22(G),level23(H),level24(I),level25(J). Model equation including all the columns and levels is as follows:

$$Y = b_0 + b_Z + b_{X_B}B + b_{X_C}C + b_{X_D}D + b_{X_E}E + b_{Y_G}G + b_{Y_H}H + b_{Y_I}I + b_{Y_J}J \quad (3.1)$$

In the equation, b_0 is the intercept. It also represents the expected value of target variable when categorical variable X is level1(A) and categorical variable Y is also level1(F).

The production of simulated data encompasses three distinct scenario types as below.

3.0.1.1 Varying true coefficient of one level of one dummy variable

The term true coefficient refers to the coefficient of an independent variable in the linear equation model when the residual is zero. For this particular scenario, we modify the true coefficient for one level of one categorical variable. The categorical variable being referred to is X, precisely its fifth level (E). The true coefficient values are fixed for other coefficients and can be seen in Table 3.1. Five datasets are generated, where K, the true coefficient of E of X, is varied from 1 through 5. True coefficient values:

Coefficient	Coefficient Value	Coefficient	Coefficient Value
A	1	F	1
B	1	G	3
C	2	H	1
D	2	I	4
E	k	J	6

Table 3.1: True Coefficients

The residuals for this dataset are normally distributed with mean 0 and variance 0.1. This scenario comprises five distinct datasets, where the true coefficient denoted by K (true coefficient value for category variable X, level5 = E) varies between 1 and 5. Sample size is set to 100.

3.0.1.2 Varying the sample size

In this scenario, four distinct datasets are generated, each with sample sizes of 25, 50, 100, and 500, respectively.K is set to 1. The variance of residual is normally distributed with mean 0 and variance 2.0.

3.0.1.3 Varying variance of residual

This scenario encompasses five distinct datasets, wherein the residual variances are set to 0.1, 0.2, 0.3, 0.7, 0.9, 1, and 20 for each dataset, respectively. K is set to 1 and Sample size=100.

3.0.2 Collinearity and least square method

The simulated data with two categorical variables has an equal number of rows per level. The level, for example A means this level belongs to categorical variable X, and is the first level, similarly B, belongs to categorical variable X, and is the second level. The same example is explained in Table 3.2.

Target	Z	X	Y
4.59	1.44	A	F
27.21	11.03	A	F
2.722	-2.172	D	I
28.48	10.79	A	F
7.610	1.803	B	G
-13.33	-7.135	A	F

Table 3.2: Sample simulated data(with collinearity)

The dataset generated suffered from collinearity. Example for the same is explained in Table 3.2. All the rows with a A for X also have a F for Y. Because of this issue, a unique least square is not possible. Multiple best-fit lines can be possible in this solution as the independent variables are related. The problem is solved by keeping the categorical variable X as is but modifying Y. Per row, the categorical variable Y's value is assigned randomly, keeping the total rows per level the same. The algorithm to assign a random value to categorical variable Y requires the total sample set to be divisible by five so that the per row per level remains the same. Table 3.3 showcases a snapshot of a dataset without collinearity.

Target	Z	X	Y
4.59	1.44	A	C
27.21	11.03	A	D
2.722	-2.172	D	J
28.48	10.79	A	J
7.610	1.803	B	B
-13.33	-7.135	A	F

Table 3.3: Sample simulated data (without collinearity)

3.0.3 Use of Linear regression

Linear regression is a supervised machine learning model that establishes a linear relationship between predictor and target variables. Linear regression can be used to understand the relationship between independent and dependent variables. Linear regression assumes that residuals (difference between actual and predicted values) are normally distributed, and have homoscedasticity. In this project, the next step is to apply the evolutionary algorithm once the simulated data is ready.

The code discusses two models: the before model and the after model. The before model includes coefficients for all levels without any replacements, while the after model includes any replaced coefficients.

3.0.4 Evolutionary Algorithm

Evolutionary Algorithm works on the raw population first. It retains the best features out of the initial population referred to as the (I) population. These best features are used further to generate a new (II) population. The generation after II generation is called the new population now, represented by (III). From population(III), the best solution or feature is chosen.

For example, the game Wordle aims to guess a word by guessing one letter at a time. The game starts with a word that the player is not aware of. Let us suppose the word is PEARL. The player will guess the first letter in the first pass. He will guess any one of the 26 alphabets; if correct, he moves to the second letter. As the letters can repeat, hence, the total complexity of this process is 26*26*26*26*26, as each letter has 26 possible values.

The algorithm examines two levels of a categorical variable and calculates the absolute difference between the coefficient values of the two levels. It then creates two models: a "before" model that includes both coefficients and an "after" model where it replaces the coefficient of one level with that of the other level, effectively reducing the number of coefficients in the model.

For this project, the simulated datasets are fed into the developed algorithm. After the algorithm reads the data, a linear regression is performed to create the before model, which includes all the independent variables. The coefficients from this model are extracted and stored for further evaluation. The categorical variable names are X and Y. All the levels of the categorical variables are compared to each other. The number of combinations to compare are ${}^{N}C_{2}$. Each pair has two corresponding coefficient values. A new set of values are created by subtracting both the coefficients and considering only the absolute value (ignoring the sign).

1) All the distinct combinations are considered. Let the cofficients be defined by A1,A2,A3,A4,A5. Here, these values are for five coefficients of five levels of categorical variable X.

$$Y = A1 + A2X_B + A3X_C + A4X_D + A5X_E$$

The distinct combinations will be (A1,A2),(A1,A3),(A1,A4),(A1,A5),(A2,A3),(A2,A4),(A2,A5), (A3,A4),(A3,A5) and (A4,A5). Table 3.4 represents the each pair and subtracted value.

First coefficient	Second coefficient	absolute difference
A1=0.3	A2=1.2	0.9
A1=0.3	A3=0.2	0.1
A1=0.3	A4=0.1	0.2
A1=0.3	A5=0.002	0.298
A2=0.3	A3=0.2	0.1
A2=0.3	A4=0.1	0.2
A2=0.3	A5=0.002	0.298
A3=0.3	A4=0.1	0.2
A3=0.3	A5=0.002	0.298
A4=0.1	A5=0.002	0.098

Table 3.4: Example coefficient values

2) Once we have obtained the absolute values, only the values less or equal than the threshold (set at 0.1) are considered for merging. Therefore, only the combination (A4,A5) will be considered for further evaluation.

3) The pair (A4,A5) is saved and proceed to the next step, where a model is generated using both coefficients of A4 and A5. This model is called the before model.In the after model, the coefficient of A4 is replaced with the coefficient of A5.
4) Using AIC, before model and after model are compared. If the after model has a better value for AIC, then that model is used for next pair of levels.
5) Once the evaluation for X is completed, the modified model is used for evaluating Y.

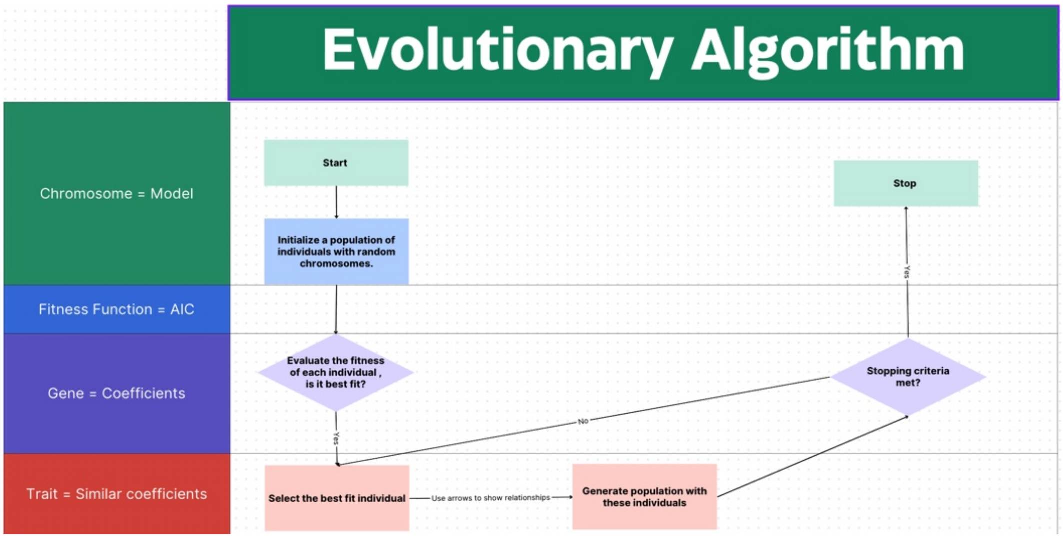
Evolutionary Algorithm
Chromosome = Model
Fitness Function = AIC
Gene = Coefficients
Trait = Similar coefficients
Start
Initialize a population of individuals with random chromosomes.
Evaluate the fitness of each individual , is it best fit?
Yes
Select the best fit individual
Use arrows to show relationships
Generate population with these individuals
Stopping criteria met?
No
Yes
Stop

Figure 3.1: An evolutionary algorithm

3.0.4.1 Census data analysis

The census data includes thirteen columns, with the number of hours worked by each individual being the response variable. The remaining columns represent independent variables such as age, workclass, education, marital status, profession, relationship, race, sex, capital gain, capital loss, native country, and salary. Among these variables, two categorical variables have been selected for analysis: marital status and profession.

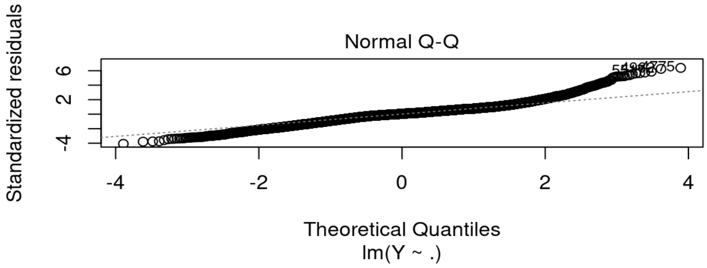

Figure 3.2: Coefficient (Intercept and X1) Vs K

Figure 3.2 shows that the census data is distributed normally.

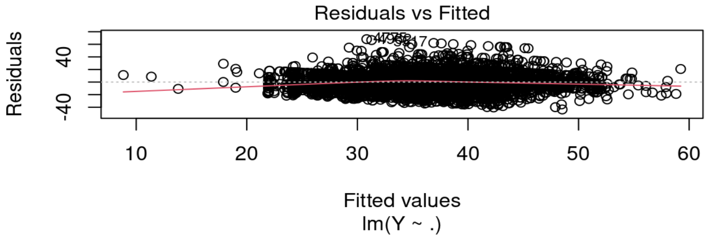

Figure 3.3: Coefficient (Intercept and X1) Vs K

Based on Figure 3.3, it can be inferred that the residuals are uniformly distributed and exhibit a linear pattern.

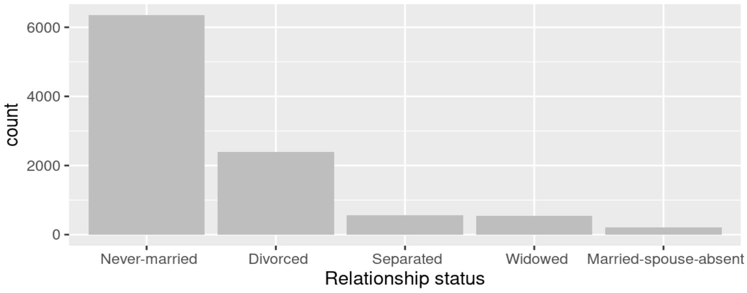

Figure 3.4: Coefficient (Intercept and Z) Vs K

Figure 3.4 displays the various levels of the relationship status variable, which is a categorical variable with five unique values.

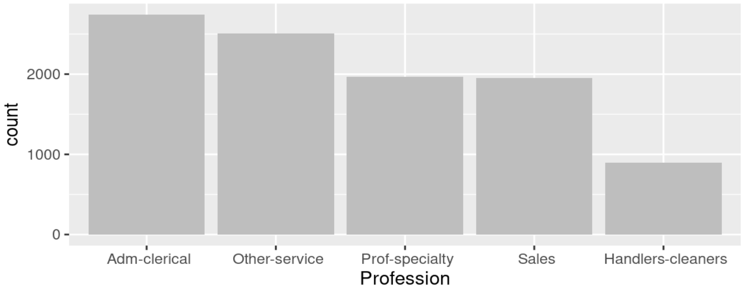

Figure 3.5: Coefficient (Intercept and Z) Vs K

Figure 3.5 displays the various levels of the profession variable for an individual, which is also a categorical variable with five distinct values.

Chapter 4

Results

4.0.1 Varying X variable's Coefficient for E

Upon varying X's level5 (E)'s coefficient value, As in Figure 4.1, it was seen that there is no change in the other coefficients as lines representing true coefficients(solid line in black and purple) and the estimated values of coefficients (represented by dahsed black and purple lines) from other datasets, for the same coefficient (intercept and coefficient of Z) are almost symmetrical.

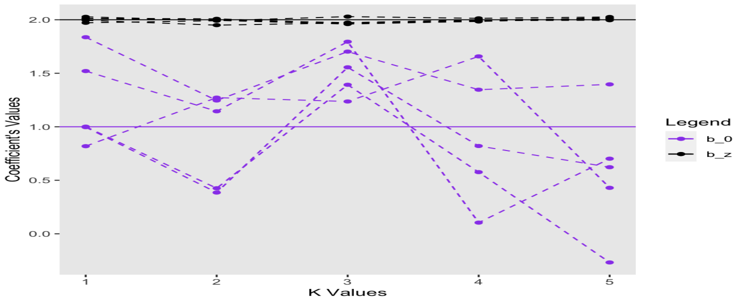

Figure 4.1: Coefficient (Intercept and Z) Vs K

In Figure 4.1, The dashed line represents the estimated coefficient value of Intercept and Z (the continuous variable) vs. the coefficient value of X variable's 5th category (E) represented as K. The solid line represents the true coefficient value. Black lines represent values for continuous variable Z and purple for the Intercept. The Intercept as by definition, also represents the reference levels for categorical variables X and Y, i.e., the level1(E) of the X variable and the level1(F) of Y. Table 4.1 contains the true coefficient and estimated coefficient values for the Intercept. Similarly, Table 4.2 contains the true coefficient and estimated coefficient values for Z. The differences between true and estimated coefficients are very minute.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	Intercept	1	0.99
Dataset2	Intercept	1	0.97
Dataset3	Intercept	1	1.02
Dataset4	Intercept	1	0.98
Dataset5	Intercept	1	0.93

Table 4.1: Comparison between true and estimated coefficients for Intercept

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	Z	2	1.99
Dataset2	Z	2	1.99
Dataset3	Z	2	2.00
Dataset4	Z	2	2.01
Dataset5	Z	2	2.00

Table 4.2: Comparison between true and estimated coefficients for Z

In Figure 4.2, The X and Y axes seem the same here, where the Y axis represents the coefficient value of the level5 of variable Z, and the X axis represents the K value set. These values are very close but different. The table contains the true coefficient and estimated coefficient values for Z variable's, 5th level(E). Table 4.3 lists the values of K and estimated coefficients of E of X categorical variable.

Dataset	Coefficient of	K	estimated coefficient
Dataset1	E	1	1.02
Dataset2	E	2	1.99
Dataset3	E	3	2.95
Dataset4	E	4	3.96
Dataset5	E	5	5.04

Table 4.3: Comparison between K and estimated coefficients for X variable's 5th level (E)

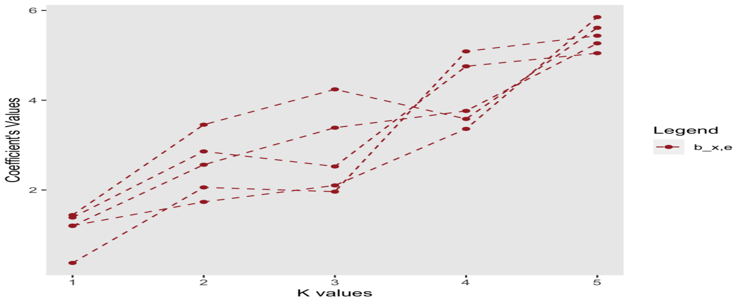

Figure 4.2: Coefficient of X variable's 5th level(E) Vs K

In Figure 4.3, The solid line represents the true coefficient value. Blue lines represent values for the X2 variable's 2nd category (B); green lines represent values for the X2 variable's 4th category(D). Red lines represent values for the X variable's 3rd category(C) (to note that the dashed red line is under the dashed green line, which means the true coefficient of the 3rd and 4th categories is the same). Table 4.4 lists the values for this scenario.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	B	1	1.02
Dataset2	B	1	0.99
Dataset3	B	1	0.95
Dataset4	B	1	1.00
Dataset5	B	1	1.04
Dataset1	C	2	2.01
Dataset2	C	2	1.98
Dataset3	C	2	2.01
Dataset4	C	2	1.99
Dataset5	C	2	2.00
Dataset1	D	2	2.02
Dataset2	D	2	1.97
Dataset3	D	2	1.98
Dataset4	D	2	1.98
Dataset5	D	2	2.01

Table 4.4: Comparison between true and estimated coefficients for X variable's 2nd, 3rd and 4th levels (B,C,D)

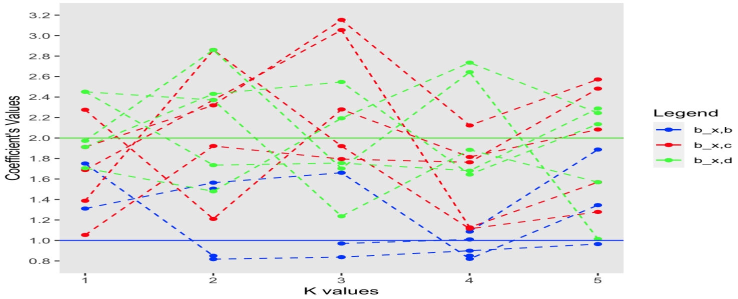

Figure 4.3: Coefficient of X variable's 2nd,3rd and 4th (B,C,D) Vs K

In Figure 4.4, The solid line represents the true coefficient value. Blue lines represent values for the Y variable's 2nd category(G), red lines represent values for the Y variable's 3rd category(H), green lines represent values for the Y variable's 4th category(I), and purple lines represent values for the Y variable's 5th category(J). Table 4.5 lists the exact values for this scenario.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	G	3	2.97
Dataset2	G	3	3.07
Dataset3	G	3	2.98
Dataset4	G	3	3.04
Dataset5	G	3	3.02
Dataset1	H	1	0.95
Dataset2	H	1	1.02
Dataset3	H	1	0.97
Dataset4	H	1	1.04
Dataset5	H	1	1.04
Dataset1	I	4	4.01
Dataset2	I	4	4.08
Dataset3	I	4	3.97
Dataset4	I	4	4.04
Dataset5	I	4	4.01
Dataset1	J	6	5.95
Dataset2	J	6	6.05
Dataset3	J	6	5.98
Dataset4	J	6	6.00
Dataset5	J	6	6.02

Table 4.5: Comparison between true and estimated coefficients for Y variable's 2nd, 3rd, 4th and 5th levels (G,H,I,J)

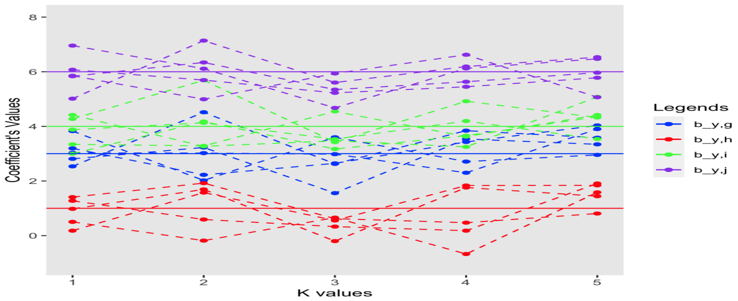

Figure 4.4: Coefficient of Y variable's 2nd,3rd, 4th and 5th (G,H,I,J) Vs K

It can be seen that with a change in K (the true coefficient of level5 of the X variable), there is no effect on the estimated coefficients for all the levels of both categorical variables X and Y.

Regarding improving model complexity, four coefficients were replaced with other coefficients for one dataset out of the five. Hence, the model dimensionality is reduced by four.

4.0.2 Parameters Vs sample size

In Figure 4.5, The solid line represents the true coefficient value. Black lines represent values for continuous variable Z and purple for the intercept. The intercept, by definition, also represents the reference categories for categorical variables X and Y i.e., level1(A) of X variable and level1(F) for Y. Table 4.6 and Table 4.7 lists the exact values for this scenario.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	Intercept	1	1.01
Dataset2	Intercept	1	1.01
Dataset3	Intercept	1	0.99
Dataset4	Intercept	1	0.96

Table 4.6: Comparison between true and estimated coefficients for Intercept

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	Z	2	2.00
Dataset2	Z	2	1.99
Dataset3	Z	2	2.00
Dataset4	Z	2	2.00

Table 4.7: Comparison between true and estimated coefficients for Z

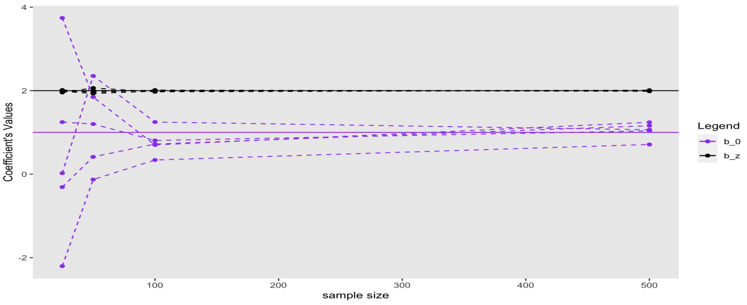

Figure 4.5: Coefficient (Intercept and Z) Vs sample size

In Figure 4.6, The dashed line represents the true coefficient value. Blue lines represent values for the X variable's 2nd category(B), green lines represent values for the X variable's 4th category, and red lines represent values for the X variable's 3rd category(C) (to note the solid red line is under the solid green line that means true coefficient of 3rd(C) and 4th category(D) is the same). Lastly, brown lines represent values for X variable's 5th category(E).Table 4.8 lists the exact values for this scenario.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	B	1	0.98
Dataset2	B	1	1.03
Dataset3	B	1	1.01
Dataset4	B	1	1.02
Dataset1	C	2	1.94
Dataset2	C	2	1.99
Dataset3	C	2	1.99
Dataset4	C	2	2.01
Dataset1	D	2	1.93
Dataset2	D	2	2.06
Dataset3	D	2	1.98
Dataset4	D	2	2.02
Dataset1	E	3	3.03
Dataset2	E	3	3.01
Dataset3	E	3	2.96
Dataset4	E	3	3.06

Table 4.8: Comparison between true and estimated coefficients for Y variable's 2nd, 3rd, 4th and 5th levels (G,H,I,J)

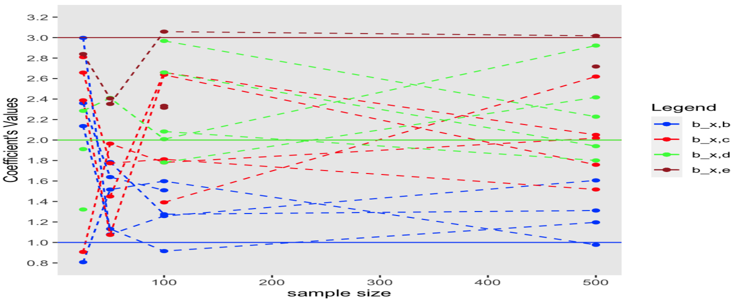

Figure 4.6: Coefficient of Y variable's 2nd,3rd, 4th and 5th (G,H,I,J) Vs sample size

In Figure 4.7, The dashed line represents the true coefficient value. Blue lines represent values for the Y variable's 2nd category, red lines represent values for the Y variable's 3rd category, green lines represent values for the Y variable's 4th category, and purple lines represent values for the Y variable's 5th category.Table 4.9 lists the exact values for this scenario.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	G	3	2.98
Dataset2	G	3	2.95
Dataset3	G	3	3.00
Dataset4	G	3	3.03
Dataset1	H	1	1.01
Dataset2	H	1	0.97
Dataset3	H	1	1.05
Dataset4	H	1	1.00
Dataset1	I	4	3.95
Dataset2	I	4	4.03
Dataset3	I	4	4.03
Dataset4	I	4	4.00
Dataset1	J	6	6.01
Dataset2	J	6	6.01
Dataset3	J	6	6.04
Dataset4	J	6	6.01

Table 4.9: Comparison between true and estimated coefficients for Y variable's 2nd, 3rd, 4th and 5th levels

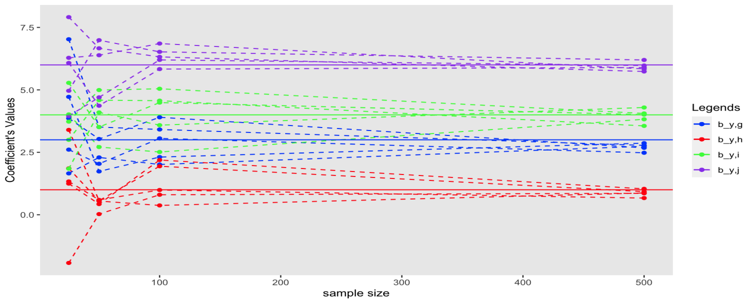

Figure 4.7: Coefficient of Y variable's 2nd,3rd, 4th and 5th Vs sample size

It can be seen that with a change in sample size there is no affect on the estimated coefficients for all the levels of both categorical variables X and Y.

Regarding improving model complexity, three coefficients were replaced with other coefficients for one dataset out of the four.Hence, the model dimensionality is reduced by three.

4.0.3 Parameters Vs variance of residuals

In Figure 4.8, The dashed line represents the true coefficient value. Black lines represent values for continuous variable Z and purple for the intercept. The intercept, by definition, also represents the reference categories for categorical variables X and Y, i.e., level1(A) of X variable and level1(F) for Y. Table 4.10 and Table 4.11 lists the exact values for this scenario.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	Intercept	1	1.00
Dataset2	Intercept	1	0.91
Dataset3	Intercept	1	1.13
Dataset4	Intercept	1	0.80
Dataset4	Intercept	1	0.366
Dataset5	Intercept	1	0.96

Table 4.10: Comparison between true and estimated coefficients for Intercept

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	Z	2	1.99
Dataset2	Z	2	1.99
Dataset3	Z	2	2.01
Dataset4	Z	2	2.00
Dataset5	Z	2	2.01
Dataset6	Z	2	2.27

Table 4.11: Comparison between true and estimated coefficients for Z

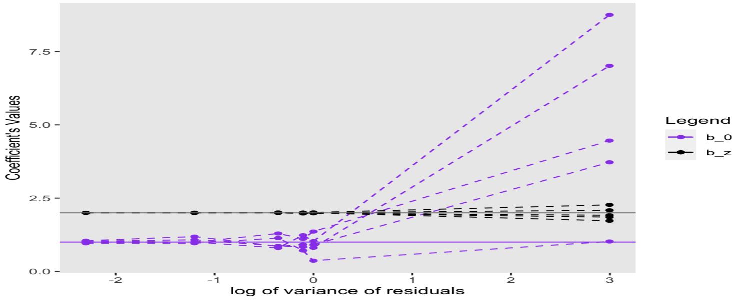

Figure 4.8: Coefficient (Intercept and Z) Vs variance of residuals

In Figure 4.9, The dashed line represents the true coefficient value. Blue lines represent values for the X variable's 2nd category; green lines represent values for the X variable's 4th category; red lines represent values for the X variable's 3rd category (to note the solid red line is under the solid green line that means true coefficient of 3rd and 4th category is the same and lastly brown lines represent values for X variable's 5th category. Table 4.12 lists the actual values of coefficients.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	B	1	1.03
Dataset2	B	1	0.97
Dataset3	B	1	0.68
Dataset4	B	1	1.03
Dataset5	B	1	1.44
Dataset6	B	1	4.11
Dataset1	C	2	2.01
Dataset2	C	2	1.88
Dataset3	C	2	2.09
Dataset4	C	2	1.91
Dataset5	C	2	2.04
Dataset6	C	2	1.02
Dataset1	D	2	2.02
Dataset2	D	2	1.94
Dataset3	D	2	1.91
Dataset4	D	2	1.85
Dataset5	D	2	2.14
Dataset6	D	2	1.12
Dataset1	E	3	3.01
Dataset2	E	3	2.96
Dataset3	E	3	2.68
Dataset4	E	3	2.71
Dataset5	E	3	3.42
Dataset6	E	2	7.4

Table 4.12: Comparison between true and estimated coefficients for X variable's 2nd, 3rd, 4th and 5th levels (B,C,D,E)

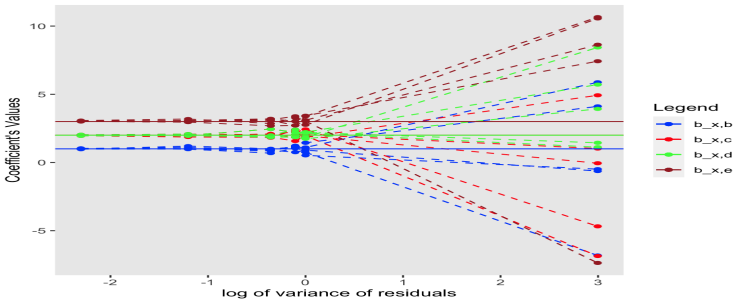

Figure 4.9: Coefficient of X variable's 2nd,3rd and 4th Vs variance of residuals

In Figure 4.10, The dashed line represents the true coefficient value. Blue lines represent values for the Y variable's 2nd category, red lines represent values for the Y variable's 3rd category, green lines represent values for the Y variable's 4th category, and purple lines represent values for the Y variable's 5th category. Table 4.13 lists the actual values of coefficients.

Dataset	Coefficient of	True coefficient	estimated coefficient
Dataset1	G	3	2.97
Dataset2	G	3	3.22
Dataset3	G	3	2.86
Dataset4	G	3	3.37
Dataset5	G	3	3.29
Dataset6	G	3	1.11
Dataset1	H	1	0.95
Dataset2	H	1	1.08
Dataset3	H	1	0.84
Dataset4	H	1	1.37
Dataset5	H	1	1.42
Dataset6	H	1	5.09
Dataset1	I	4	4.01
Dataset2	I	4	4.25
Dataset3	I	4	3.79
Dataset4	I	4	4.41
Dataset5	I	4	4.16
Dataset6	I	4	-6.02
Dataset1	J	6	5.95
Dataset2	J	6	6.17
Dataset3	J	6	5.86
Dataset4	J	6	6.08
Dataset5	J	6	6.26
Dataset6	J	6	15.5

Table 4.13: Comparison between true and estimated coefficients for Y variable's 2nd, 3rd, 4th and 5th levels(G,H,I,J)

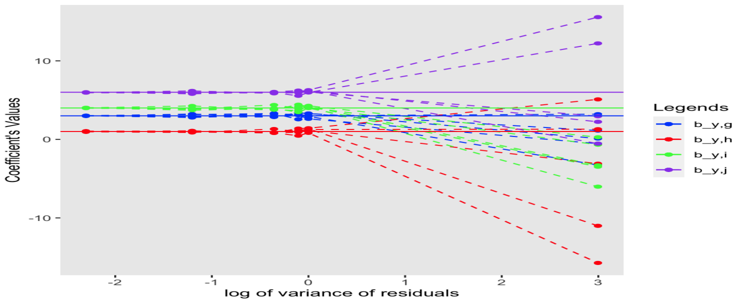

Figure 4.10: Coefficient of Y variable's 2nd,3rd, 4th and 5th Vs variance of residuals

For the intercept, the estimated coefficient value fluctuated from the true value of 1 to 0.366. This is a difference of almost 0.6. It means that for a population, the average value of the target variable is 0.366 when the X variable is level1(A), and the Y variable is also level1(A) when the average value of the target variable is 1 when X and Y variables are A and F respectively.

In the case of continuous variable Z, there is little change, and the model can be considered an average model as the fluctuation from 2 to 2.27 is not huge.

For the X variable, with the increase in the variance of residuals, there is a vast deflection in the estimated coefficient's value. The most change for level1 is of 3 units. The most deflection is for E, with a deflection of almost 5.5 units from the true coefficient value.

For Y variable, the major deflection from true value and estimated value is seen for level4(I) and level5(J). For level4 the deflection is almost 10 units. For level5(J), the deflection is of 8.5.

The model becomes a poorly estimated model with large residuals.

Regarding improving model complexity, one coefficient was replaced with other coefficients for one dataset out of the six (the dataset with larget variance of residual). Hence, the model dimensionality is reduced by one.

4.0.4 Census Data results

The analysis focuses on two categorical variables: marital status and profession. The original dataset includes options for marital status such as never married, widowed, divorced, separated, and married but with a spouse, not in contact. The variable profession has possible values, such as clerks, cleaners, professors, salespeople, and other service providers. The response variable is the number of hours an individual work. The algorithm is designed to determine if all categories' weight is necessary to predict the response variable. The algorithm indicates that the category "married but with a spouse, not in contact" is insignificant in predicting the response variable and is therefore combined with "never married." This suggests that the weight of "married but with a spouse not in contact" does not significantly contribute to the model's performance. However, all categories for the profession variable are essential in determining the hours worked.

As here, the model is getting rid of one of the parameters out of almost 20 in the model. Similarly, the algorithm can be used on other categorical variables present in the dataset.

Chapter 5

Discussion

5.0.1 Selecting the most important categorical variable first

The evolutionary algorithm employed in the project aims to reduce the model's dimensionality. The algorithm starts by processing the first column encountered, denoted as X, and operates on its levels to combine two or more categories, if doing so improves the AIC score. If combining or replacing of coeffecients fails to improve the AIC score, the levels remain separate. Once all levels of X are checked, the dataset's state is then moved to the next column, Y, and each of its levels is examined for potential merging.

Modifying X works on the before or raw model, whereas Y is only modified after model is already modified due to X. To improve the algorithm, it would be beneficial to select the most critical feature based on feature selection methods or field knowledge and prioritize modifying the dataset based on that feature. This way, the more important feature is processed first, ensuring that the most necessary modifications are made to the model first.

5.0.2 Comparing Lasso to Evolutionary Algorithm

As we know from the lasso algorithm, it shrinks the coefficient of irrelevant features to zero. If lasso is applied to the simulated dataset, the coefficients of some of the variables shrink to zero. The impact of these variables is entirely nullified. This is not the case with this algorithm. For a particular dataset, the coefficients of two levels, B and C, are replaced with E and D, respectively. For the same dataset, Y's all levels remained as such. Applying lasso to this dataset with unmerged levels is a case of all or nothing. The lasso model will either include all five levels and their respective penalization or exclude them altogether. Lasso does not allow us to compare the different levels and their effect on the predictor variable.

5.0.3 Algorithm Limitation and big data application

The current algorithm is designed to function with a limited scope of two categorical variables, each having no more than five levels. However, its potential can be expanded to cater to additional categorical variables and levels. It is crucial to distinguish this algorithm from a linear model as the latter treat categorical variables with numerical values as continuous variables without separate coefficients. In the case of numerical values embedded within categorical variables, such as 0, 1, 2, and 3, it becomes necessary to factor the categorical variable internally before utilizing linear regression. This algorithm considers any type of value as levels for a categorical variable.

When working with compute-intensive algorithms, the matter of complexity becomes paramount. The algorithm's space and time complexity can pose significant challenges when dealing with large datasets. Several approaches can be employed to mitigate this, such as discarding intermediate results that are no longer necessary after the model's dimensions have been reduced. However, this particular algorithm retains intermediate results to enable backtracking, which can also contribute to its complexity.

5.0.4 Use of CV to find the threshold

For the current algorithm, threshold for coefficient's absolute difference is set to 0.1. The most optimal value for the threshold can be determined in the developed algorithm by setting it as the CV.

Bibliography

[1] H. Wainer, "Estimating coefficients in linear models: It don't make no nevermind." *Psychological Bulletin*, vol. 83, no. 2, p. 213, 1976.

[2] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. b*, vol. 4, pp. 51–62, 2017.

[3] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[4] H. Abdi *et al.*, "The method of least squares," *Encyclopedia of measurement and statistics*, vol. 1, pp. 530–532, 2007.

[5] J. E. Cavanaugh and A. A. Neath, "The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 3, p. e1460, 2019.

[6] H. Alkharusi, "Categorical variables in regression analysis: A comparison of dummy and effect coding," *International Journal of Education*, vol. 4, no. 2, p. 202, 2012.

[7] B. Ratner, "The correlation coefficient: Its values range between+ 1/-

1, or do they?" *Journal of targeting, measurement and analysis for marketing*, vol. 17, no. 2, pp. 139–142, 2009.

[8] J. W. Osborne and E. Waters, "Four assumptions of multiple regression that researchers should always test," *Practical assessment, research, and evaluation*, vol. 8, no. 1, p. 2, 2002.

[9] M. R. Shaikh and J. Beyene, "Statistical models and computational algorithms for discovering relationships in microbiome data," *Statistical Applications in Genetics and Molecular Biology*, vol. 16, no. 1, pp. 1–12, 2017.

[10] T. Bäck and H.-P. Schwefel, "An overview of evolutionary algorithms for parameter optimization," *Evolutionary computation*, vol. 1, no. 1, pp. 1–23, 1993.

[11] L. Van Der Maaten, E. Postma, J. Van den Herik *et al.*, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[12] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[13] L. I. Smith, "A tutorial on principal components analysis," 2002.