LINEAR REGRESSION MODEL

*36106 Machine Learning and Algorithm Aplication*

|  |  |
| --- | --- |
| Student Name: | Student ID: |
|  |  |
|  |  |
| Bui The Hai | 24683423 |
|  |  |
|  |  |

# Introduction, Problem Statement and Research Questions

Introduction

During this project, I will perform linear regression models on Cancer Death Rate of US counties. The dataset has one independent variable TARGET deathrate and 34 independent variables. The project is divided into three main parts. The first part is the performance of univariate linear regression on two different variables. The second part will perform multivariate linear regression on various variables. The final one will be the multivariate linear regression with feature engineering on variables.

Problem Statement and Questions

# We will analyze the various factors for cancer death rate and by using linear regression models. In this part, I will propose two hypotheses:

# Median of Income per county has a reverse relationship with TargetDeathRate (Mean per capita (100,000) cancer mortalities)

# Percent of populace in poverty (povertyPercent) has an positive relationship with TargetDeathRate (Mean per capita (100,000) cancer mortalities)

# Business understanding:

# Cancer is one of the biggest problems that human beings face in the 21st century. There are many factors that can affects the death rates of cancers. This essay will focus on analyzing factors having big impacts on cancer death rate by using multi linear regression method.

# Data preparation:

# For data preparation part, before running the model, I will perform checking the datasets.

# 

# Figure 1. Test Dataset

# 

# Figure 2. Train Dataset

# We have two datasets named train dataset and test dataset. In the data frames, we have 29 float variables, 4 int variables and 2 object variables.

# I will combine this set into one data set, then split it to training and validation set by a 40-60 ratio.

# Modeling:

# Select data:

# In the modelling part, I will divide it into three parts. In part one, I will perform two univariate linear regression models with the dependent variables TARGET\_deathRate, and the two independent variables are povertyPercent and medianIncome. The reason for choosing these two independent variables is that I will answer the question whether income level can affect the target death rate due to cancer.

# In part two, I will run multilinear regression models on eight independent variables, namely incidentRate, medIncome, studyPerCap, MedianAge, Geography, AvgHouseholdSize, PctNoHS18\_24, PctHS18\_24, BirthRate. The reason I choose these variables because my aims is to test the impact on the target death rate of various socioeconomic factors

# In part C, I will use log transformation to above variables in Part B before running a regression model. The reason why I perform log transformation is that the data is not normally distributed. As a results, logarithmic transformation will transform the data to be normally distributed.

# For example, the medIncome variable is skewed, so it is better to perform log transformation on this variable

# Clean data:

# In the cleanning part, I will examine the data to see whether it has any missing values in the dataframe. The dataframes do have many missing values so I will fill the missing values with its mean.

# A screenshot of a data table Description automatically generated

|  |  |
| --- | --- |
| Missing value | No of rows |
| PctSomeCol18\_24 | 2285 |
| PctEmployed16\_Over | 152 |
| PctPrivateCoverageAlone | 609 |

# The variable ***PctSomeCol18\_24*** miss 2285 out of 3047, which should be exclude from the model. ***PctEmployed16\_Over and PctPrivateCoverageAlone*** variables will be imputed by its mean value.

# The first model: Univariate Linear regression model:

The goal of this experiment is to examine the relationship between wealth and cancer death rate ***(TARGET\_deathRate).*** Therefore, I will choose two variables, namely ***medIncome*** and ***povertyPercent*** as independent variables and train two univariate linear regression models.

# 

# ***TARGET\_deathRate and medianIncome***

# 

# ***TARGET\_deathRate and povertyPercent***

# The model clearly show an reverse relationship between TARGET\_deathRate and medianIncome and an invererse realtionship with povertypercent

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MSE of baseline | MSE of training | MSE of testing | Coefficients |
| Regression models for ***TargetDeathRate*** and ***povertyPercent*** | 768.13 | 624.95 | 640.07 | 1.86 |
| Regression models for ***TargetDeathRate*** and  ***medIncome*** | 768.13 | 631.07 | 618.54 | -0.00096 |

# Comparing the result to the baseline, it is indicated that the mode performs better. The coefficient highlights the reverse relationship between TargetDeathRate and medIncome. On the other hand,it is indicated that povertyPercent illustates the inverse relationship between povertyPercent and TargetDeathRate.

# According to the results of the model, one can conclude that the wealth level of a person can have a an impact on the target death rates due to cancer. This relationship may have many profound and different meanings in many areas of our society. For example, from the government viewpoint, this experiment emphasizes the bad influence of income parity in US society, and they need to find a way to tackle these problems. However, dealing with cancer needs the joint efforts of everyone in our society. The participation of entrepreneurs is a good solution when they can balance social responsibility and profit maximization. For example, they can get an edge on their rivalry in business areas like insurance.

# The second model: Multivariate Linear regression model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MSE (baseline) | MAE (baseline) | MSE  (training set) | MAE  (training set) | MAE (validation set) | MSE (validtion set) |
| Model 3: Multivariate linear regression | 585.27 | 19.42 | 377.37 | 15.05 | 347.7 | 14.5 |

# MSE and MAE table

|  |  |
| --- | --- |
| ***Intercept*** | 177.70 |
| ***incidenceRate*** | 9.18 |
| ***medIncome*** | -9.58 |
| ***studyPerCap*** | -0.94 |
| ***MedianAge*** | -0.45 |
| ***AvgHouseholdSize*** | 1.11 |
| ***PctNoHS18\_24*** | 1.03 |
| ***PctHS18\_24*** | 4.19 |
| ***BirthRate*** | -1.73 |

# Coefficient table

|  |  |
| --- | --- |
| A graph with red dots  Description automatically generated | A graph with red dots  Description automatically generated |
| Training set | Testing Set |

# In the second model, I perform multivariate linear regression model to illustrate the relationship between the cancer death rate and all other socioeconomic factors.

# Firstly, comparing the results with the baseline shows that the model performs better. Secondly, the coefficient table shows that there is indeed a negative relationship between median income and cancer death rate, which is aligned with finding on the first experiment. So, areas with higher median incomes tend to have lower cancer death rates, which aligns with what we would typically expect in terms of access to healthcare, better lifestyle choices, and other socioeconomic factors associated with higher income levels.

# Secondly, higher study per capita is associated with lower cancer death rates. This suggests that areas with more educational resources or research tend to have lower cancer death rates.

# Thirdly, there appears to be a positive relationship between the incidence rate and cancer death rate. This suggests that areas with higher incidence rates tend to have higher cancer death rates.

# The third model: Multivariate Linear regression model:

# In this part I will perform the log transformation of the variables. The reason is that independent variables are skewed. As a result, the logarithm transformation will bring the variables close to normal distribution. Outliers will also be removed to make the model symmetric.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MSE of baseline | MAE of baseline | MSE of model on training set | MAE of model on training set | MSE of model on testing set | MAE of model on testing set |
| Model 4: Multivariate linear regression | 527.63 | 18.36 | 300.88 | 13.46 | 300.34 | 13.32 |
| Model 5:  Multivariate linear regression  (Using Lasso model) |  |  | 306.14 | 13.62 | 303.12 | 13.43 |
| Model 6:  Multivariate linear regression  (Using Ridge model) |  |  | 300.88 | 13.46 | 300.32 | 13.31 |
| Model 7:  Multivariate linear regression  (Using Elastic Net model) |  |  | 332.6 | 14.33 | 320.26 | 14.00 |

# Lasso and Ridge regularization will be performed on the data sets. From this comparison, it appears that the Multivariate Linear Regression model without regularization (baseline) has the highest MSE and MAE on the training set, indicating overfitting. However, on the testing set, the Ridge model and the baseline model have the lowest MSE and MAE, suggesting better generalization performance. The Elastic Net model has the highest MSE and MAE on both training and testing sets, indicating it might not be the most suitable choice for this dataset.

# Overall, the Ridge model seems to strike a good balance between bias and variance, offering decent performance on both the training and testing sets. Below is the illustration of Ridge model.

# A graph with blue and orange squares

|  |  |
| --- | --- |
| A graph with red dots  Description automatically generated | A graph with red dots  Description automatically generated |
| Training set | Testing set |

The model results have many implications. Median income has the most influence on cancer death rate predictions. This suggests that socio-economic factors, particularly income levels, play a crucial role in determining cancer mortality rates, as higher median incomes are associated with lower cancer death rates.

AvgHouseholdSize also has a significant influence on cancer death rate predictions. This suggests that socio-economic factors associated with larger households, such as lower income levels or limited access to healthcare, may contribute to poorer cancer outcomes.

The coefficients for PctNoHS18\_24 and PctHS18\_24 indicate that education levels among young adults correlate with cancer mortality rates. Higher education levels are associated with lower cancer death rates, emphasizing the importance of education in promoting health literacy, awareness of preventive measures, and adherence to healthcare recommendations.

The coefficient for studyPerCap suggests that areas with higher educational resources or research activities tend to have slightly lower cancer death rates. This implies that access to healthcare resources, including educational institutions and research facilities, can positively impact cancer outcomes by facilitating early detection, innovative treatments, and dissemination of health information.

The coefficient for BirthRate suggests that areas with higher birth rates tend to have slightly lower cancer death rates. This may reflect the age distribution of the population, with younger populations potentially having lower cancer mortality rates due to different disease prevalence or screening behaviors.

# Evaluation:

In my opinion, after training all the models, it is evident that the wealth level has a big impact on the target death rate. It is also highlighted by the effect of education level. People who got a higher education are less prone to risk of cancer death. Model 6 perform well on both training and validation with consistently MSE and MAE should be the best model.

The results will be referenced for further experiments on factors affecting cancer death rate. My suggestion is that the study should be carried out on a larger scale and in a different community to obtain a broader picture.

Cancer is a dilemmas problem in our society. Analyzing factors affecting cancer death rate will pave a way for eradicate this disease soon.

**REFERENCES**

* + - 1. Notebook design template: Anthony So
      2. Hotz, N. (2023) *What is CRISP DM?, Data Science Process Allian*ce. Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: 12 February 2024).