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.

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:

1. Median of Income per county has a reverse relationship with TargetDeathRate (Mean per capita (100,000) cancer mortalities)
2. 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. In my experiment, I will eliminate all the two object variables before performing regression models.

# 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 'PctNoHS18\_24', 'PctHS18\_24', 'PctSomeCol18\_24', 'PctBachDeg18\_24',' PctHS25\_Over',' PctBachDeg25\_Over', 'PctEmployed16\_Over', 'PctUnemployed16\_Over'. The reason I choose these variables because my aims is to test whether education level has an impact on the target death rate of people who get cancer.

# In part C, I will use log transformation to run a regression model on all variables. 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 not have many missing values so I will fill the missing values with 0.

# 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 ***medianIncome*** 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 | MAE of baseline | MSE of model | MAE of model | Coefficient | Intercept |
| Model 1: Regression models for ***TargetDeathRate*** and ***medIncome*** | 27.53 | 21.31 | 24.76 | 18.86 | -0.001 | 226.33 |
| Model 2: Regression models for ***TargetDeathRate*** and ***povertyPercent*** | 27.53 | 21.31 | 24.73 | 18.85 | 1.89 | 146.9 |

# Comparing the result to the baseline, it is indicated that the mode performs better. The coefficient highlights the reverse relationship between ***TargetDeathRate and medIncome.*** IOn 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 view point, 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 of baseline | MAE of baseline | MSE of model | MAE of model |
| Model 3: Multivariate linear regression | 758.055 | 21.31 | 641.2 | 18.2 |

# MSE and MAE table

|  |  |
| --- | --- |
| Intercept | 169.0139999 |
| PctUnemployed16\_Over | 2.279320084 |
| PctHS25\_Over | 0.644776806 |
| PctHS18\_24 | 0.151037299 |
| PctBachDeg18\_24 | 0.100588005 |
| PctSomeCol18\_24 | -0.016590731 |
| PctEmployed16\_Over | -0.133932735 |
| PctNoHS18\_24 | -0.483402997 |
| PctBachDeg25\_Over | -1.497963072 |

# Coefficient table

# In the second model, I perform multivariate linear regression model to test the impact of education on Cancer death rate

# Firstly, comparing the results with the baseline shows that the model performs better. Secondly, the coefficient table shows that unemployed people face a higher risk of cancer death. People aged 25 and over who achieve bachelor’s degrees face less risk of cancer death. The level of education will have a positive effect on the 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

I perform Lasso and Ridge regularization on the data sets. The Lasso model brings all the coefficients of independent variables to 0. However, in Ridge model, by adding some penalty, **medIncome, studyPerCap,avgAnnCount,PctWhite** show negative relationship with TargetDeathRate, while the other show a positive relationship with TargetDeathRate. It is interesting because it shows an inverse relationship between medIncome and TargetDeathRate, which is totally different with the second part of the experiment.

|  |  |
| --- | --- |
| Intercept | 169.0139999 |
| PctUnemployed16\_Over | 2.279320084 |
| PctHS25\_Over | 0.644776806 |
| PctHS18\_24 | 0.151037299 |
| PctBachDeg18\_24 | 0.100588005 |
| PctSomeCol18\_24 | -0.016590731 |
| PctEmployed16\_Over | -0.133932735 |
| PctNoHS18\_24 | -0.483402997 |
| PctBachDeg25\_Over | -1.497963072 |

The model results have many implications.it may mean that people with higher levels of income or white people are more prone to cancer. It may have impacts on treatment of cancer when the patients are white people, especially this study is carried out in the US, the country has many white people

# 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 highlighted by the effect of education level. People who got a higher education are less prone to risk of cancer death.

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.