Personal Final Report

1.Introduction

The pollution could be the huge problem for today's world. From the data set, it supposed the indoor and outdoor pollution for some country the number of fatality is over 200 per million people. This is several time as the fatality rate of automobile accident. In our project, we are trying to evaluate the severe of air pollution in a country by some non-monitor data that means that the predict variable are easy to access without using a large amount of sensors to evaluate the air quality.

We use linear regression model, this model usually work well with predict the feature of a large population . It has the disadvantage that it is a linear model, so, it requires carefully analysis and adjustment to produce a good model.

2. Algorithm and Model Basic

In the group project, I did the part of training model. I use the linear regression model to figure out the relationship between fatality rate, fuel useage and some other potential features. The mathematical interpretation of linear regression with multiple features input could be demonstrate by the linear algebra.

Regression function:

$$\hat{Y} = X\hat{\beta}$$

The $\hat{\beta}$ is a vector, The $\hat{\beta}$ contain N estimated parameters: first element is interception parameter and the rest are N-1 slope parameters, it is a function minimized the MSE.

A matrix X which is filled with 1 for the first column and rest of columns are filled with features

Of different observations. X_{ij} is filled with value of ith observation and j-1th feature. To find $\hat{\beta}$, we need to minimize function $\left(Y - X\hat{\beta}\right)^T\left(Y - X\hat{\beta}\right)$ by take derivative of $\hat{\beta}$, and find the zero

Since: $X^T X\hat{\beta} = X^T y$ then, $\hat{\beta} = (X^T X)^{-1} X^T y$

The data set features may not show the linear correlation, it could require data transformation and feature generation. For some data, the predicted feature and explanatory features does not show linear relationship.

Transformed the predicted feature with log, and the model will be a exponential decay model.

$$\hat{Y} = exp\left(X\hat{\beta}\right)$$

The generated explanatory features could be added into model by add a feature which is a function of original feature. For instance, square the original feature and add as a new column of feature for input X. Sometimes the previous feature could have interaction, a term of interaction feature could be add into input X by generating a feature from multiplying or dividing of previous features.

It is efficient to use residual to analysis the performance of the regression; the assumption of ideal linear regression is the residual for all of different feature values should have same variance and mean of 0. For regression with multiple explanatory features, plot regression line with observation is hard. Analysis of residual with each features is more feasible.

Residual: $\hat{e} = Y - \hat{Y}$ residual is the real value - predicted value

R-squared Metrics

R-squared suppose portion of predicting variable’s variation in observations could be explained by the model.

Since $MSE = \hat{e}^T \hat{e}/n$ and R-squared= $1 - (\frac{MSE}{Var(Y)})$

Hypothesis testing: F-test

To make selection of model, F-test is a method to compare two similar model. If model 1 contains all of explanatory features of model 2, and model 1 contain some more features. Then the model 1 is the full model and the model 2 is the reduced model. A F-test could identify whether the full model have advantage on prediction.

Confidence Interval

Use the regression model by inputting features, we can get an estimated mean for the predicted feature with specific explanatory features. A confidence Interval can generate a lower bound and a upper bound for this estimated mean.

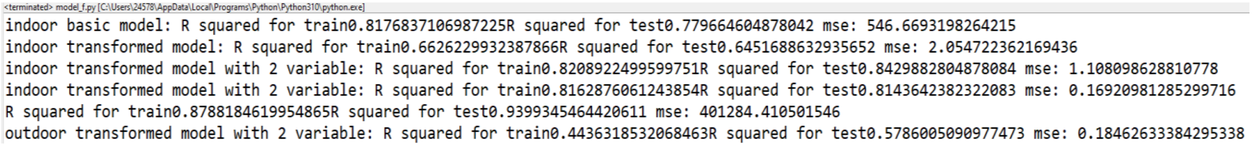
$$\hat{Y} \pm \sqrt{\left(X_0 (Z^T Z)^{-1}_{X_0}\right)}\hat{\sigma}\Phi^{-1}_{(\alpha)}$$

3.Work Description

I trained the model for estimated the fatality rate indoor and outdoor. In the coding part, I use pandas and numpy to generate data for sklearn.linear_model.LinearRegression Regressor. I use matplotlib.pyplot to generate the plot image, and use the image to evaluate the result. I use the mean_square_error form sklearn.metrics to calculate MSE and regressor’s score method to calculate the R-squared value.

I start with use the function of sklearn directly to generate the model, then I tried to start developing a module based on numpy, pandas and sklearn to process the model making. Finally I finished a python file called model maker, this file use function to analysis the input X,Y automatically, it can also adjust by a input variable to produce log transformed model, it could use a string to report finding in the model. Then I also create two auto plot function to plot residual and plot the estimated point with real point in a graph.

Fig1 The example of model_maker report



```
<terminated> model_f.py [C:\Users\24578\AppData\Local\Programs\Python\Python310\python.exe]
indoor basic model: R squared for train0.8176837106987225R squared for test0.779664604878042 mse: 546.6693198264215
indoor transformed model: R squared for train0.6626229932387866R squared for test0.6451688632935652 mse: 2.054722362169436
indoor transformed model with 2 variable: R squared for train0.8208922499599751R squared for test0.8429882804878084 mse: 1.108098628810778
indoor transformed model with 2 variable: R squared for train0.8162876061243854R squared for test0.8143642382322083 mse: 0.16920981285299716
R squared for train0.8788184619954865R squared for test0.9399345464420611 mse: 401284.410501546
outdoor transformed model with 2 variable: R squared for train0.4436318532068463R squared for test0.5786005090977473 mse: 0.18462633384295338
```

Use these function to test result of regression.

I also make a scipt called estimator for GUI, since it stores the best model I found before. When the GUI start, the model will not need to be trained again. The estimator class have the method to estimate the indoor and outdoor fetality rate with their confidence interval.

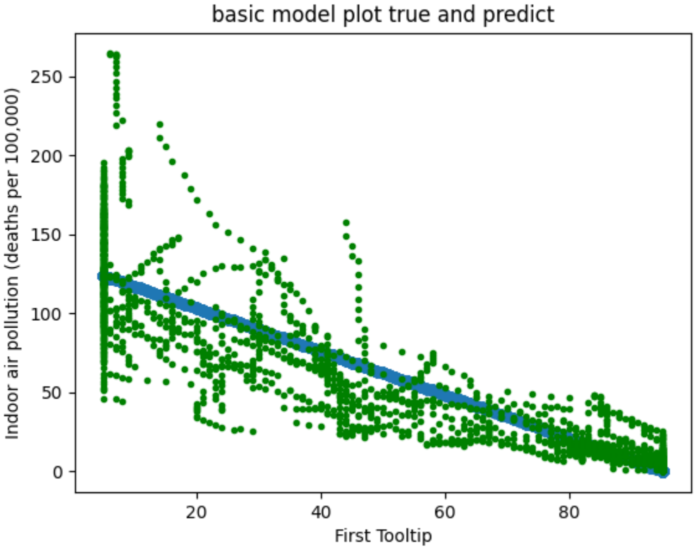
4. Results

Table 1. The best model found

	R-squared	Variables
Indoor log model	0.814	Clean fuel access rate in the population(%), Gdp per Capita(USD)
Outdoor log model	0.528	GDP/fuel(USD/kwh), GDP per capita

During training process, I received the previous EDA result from group mate and use the result to find potential features important. I generate the first basic model and plot the residual vs the portion of population access clean fuel and technology. The result does not show the constant variance. From the plot, the residual increase with the percentage of people accessing clean fuel and technology. It infers that I need to do log transformation to the fatality rate. That will infer that the model with fatality rate and percentage of population access the clean fuel and technology could be a exponential decay relationship. At this situation, the random noise term will have mean of 0 and constant noise. And the noise could be a scale noise with exponential, that means that the residual is proportional to the estimated mean. For both indoor and outdoor situation, the log transformation is required.

Fig2. The plot of basic model predict and real observations



To explore better model, I also did some preprocessing part, including feature generating and data transformation. I first want to use the medical expenditure as second variable, however, the observation of medical expenditure is limited. In the 2016 case study, I found the generally relationship that the medical expenditures and GDP are linear proportioned and the intercept is nearly zero. The R squared of this model is 0.93. Then I used the GDP and clean fuel usage rate in the main indoor model to apply estimate. After removed the extreme conditions that the usage rate is over 94. The model show well with r-squared and residuals.

The best model I found is the log model with explanatory variables: percentage of population access clean fuel and GDP per capita, it shows R-squared over 0.82 and the residual plot shows the same variance and mean of zero.

Fig3: plot the residual, x axis is the percentage of population access clean fuel

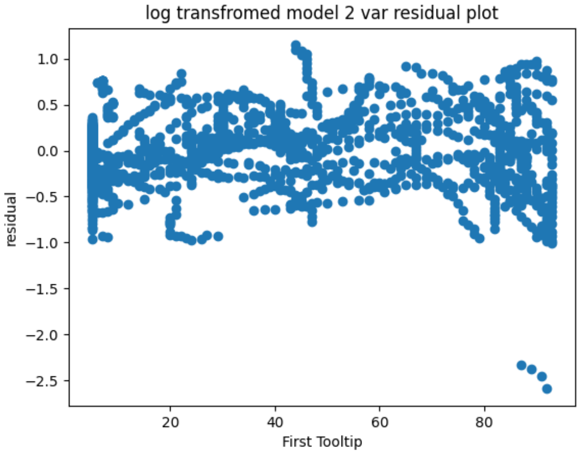


Fig4: plot the estimated result and true result, x axis is the percentage of population access clean fuel

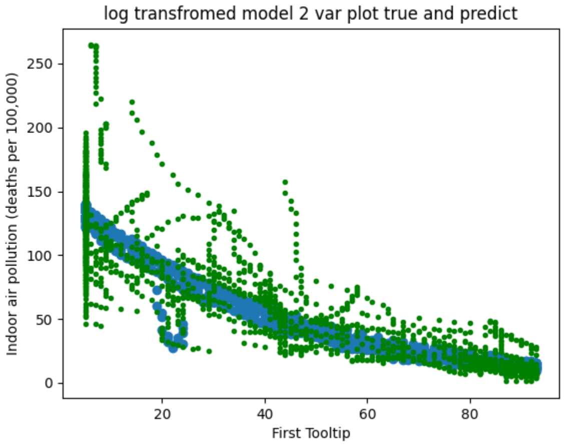
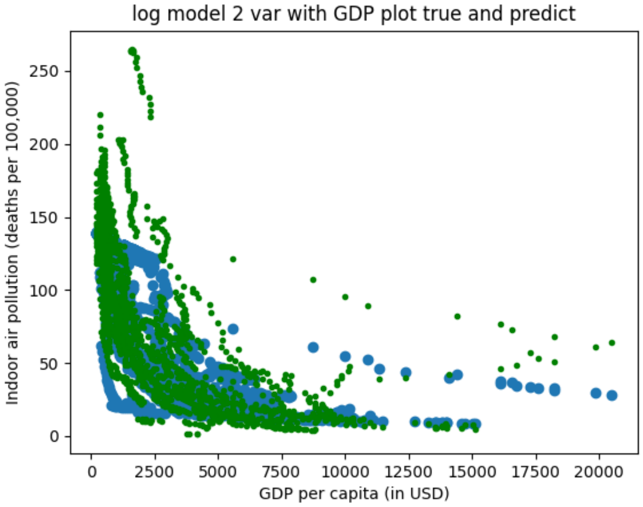


Fig5: plot the estimated result and true result, x axis is the GDP per capita



I need to use pandas map function to squared some variable and generate new features by applying function with existed features when I tried to deal with the outdoor fatality rate. Since the all of feature work not well with prediction the outdoor fatality rate. The regression model and log regression model has R-squared less than 0.1. The generated feature that average usage energy from coal per capita and The generated feature that average usage energy from coal per capita could not solve the problem. However, generated feature, GDP/fuel in the unit of USD/kwh, with feature GDP per capita make the log model with the R-squared of 0.53

Fig 6: plot the estimated result and real observations of outdoor fatality rate, x axis is GDP per capita

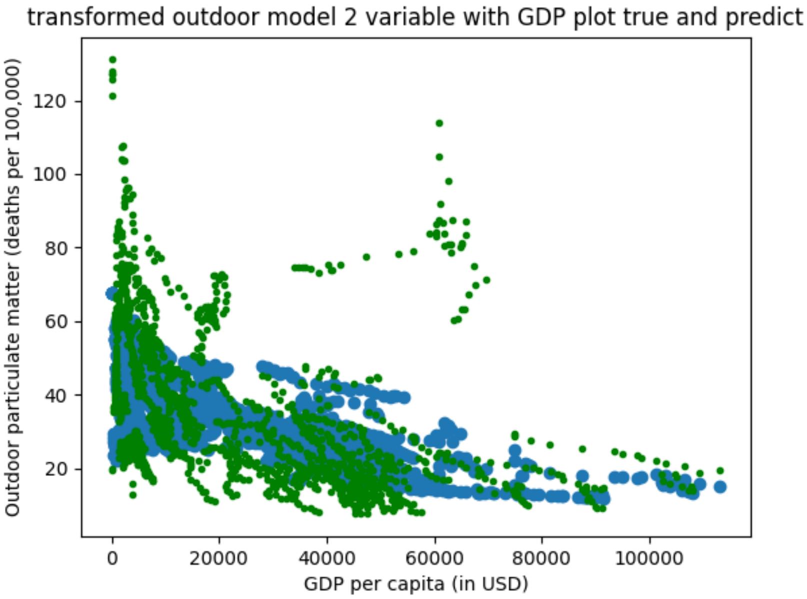
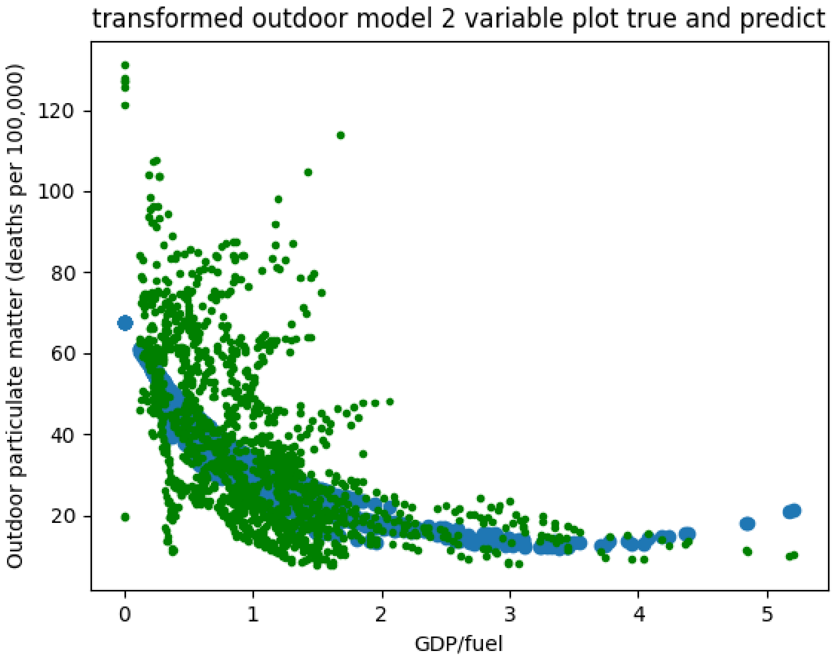


Fig 7: plot the estimated result and real observations of outdoor fatality rate, x axis is GDP/fuel



After finished the model, I work with Ali and write a estimator class to store the trained model, it could predict the indoor and outdoor fatality rate by entering the percentage of population accessing clean fuel usage, GDP/fuel ratio and GDP per capita. It also calculate the confidence interval for both estimation. It is imported in the GUI to estimate results.

5. Summary and Conclusion

From the model of indoor and outdoor, I found a strong correlation between fatality caused by indoor pollution and percentage of population accessing clean fuel usage. The usage of clean fuel in home can significantly decrease the fatality rate. For the outdoor situation, the model is not very accurate since the R-square is around 0.5. It still point out that the energy saving industries that produces more value with less energy could be good for people’s health.

Using linear regression deal with data with a lot of features could be a challenge, since it requires me to inspect the feature in detail. I need to evaluate the EDA part from group mate carefully to figure out potential candidates of good feature for estimation. Doing data transform and feature generation are also important technique in modeling. The linear regression has constraint on linearity; it cannot predict accurately without suitable preprocessing. The model still need some improvement since it predict some value with huge discrepancies. The dataset from Kaggle still have the limitation that it could not divide the regions into more subregions. It could not provide enough information to improve more accuracy. Another problem is the observation of some specific features is limited. The linear regression model’s prediction in these interval has large discrepancy. For the instance with high GDP per capita and GDP/fuel ratio, the estimation is not accurate. I need to apply some new model and preprocessing technique to increase the accuracy of extreme case.

6. Code Usage from Sources

6/(65+51+66)*100=3.2%

7. Reference

Code:

Pandas Website: https://pandas.pydata.org

Sklearn Website: https://scikit-learn.org/stable/

Mathematical Background:

Linear Model With R, by Julian J. Faraway