International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

BUILD LINEAR REGRESSION MODELS FOR INSURANCE CHARGES

Ziyi Jiang

Carnegie Mellon University

Address: 39 Lee, Irvine, CA 92620

DOI: 10.46609/IJSSER.2022.v07i02.012 URL: https://doi.org/10.46609/IJSSER.2022.v07i02.012

Received: 18 Feb. 2022 / Accepted: 27 Feb. 2022 / Published: 28 Feb. 2022

ABSTRACT

In this project, the linear regression model is performed to predict the insurance premium fees. The model is programmed by Python and the results show that it achieves pretty good accuracy, and it is easy for people to identify the important variables. Moreover, we compared Lasso and Ridge regression for further feature importance exploration.

Keywords: Linear Regression, Lasso Regression, Ridge Regression

Introduction:

Insurance is essential to any people who seek to reduce the worries of financial loss/risk after unexpected incidents happen. Under the background of Covid-19 pandemic, many things are unpredictable, which makes people care about insurance even more. In this case, insurance companies should have a model that can help them to set insurance charges individually. Hence, our goal in this research is to better understand insurance charges, and build linear regression models to predict medical expenses, which may help people who seek to buy insurance and insurance companies to better understand the determinants of insurance costs.

Data:

The dataset we use in the report is collected on a sample of 1338 insured. Each row in the data represents a person, and his age, sex, bmi, children (numbers of), smoker (whether or not), region, and charges (in dollars).

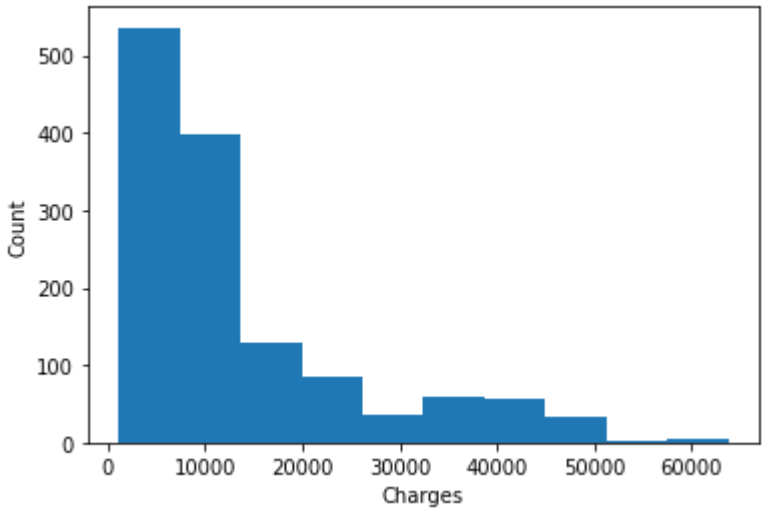
Our primary focus will be on the variable **charges**, a quantitative variable showing each individual's insurance charge. As shown in Figure 1, the distribution is unimodal and skewed to

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

the right (less typical on higher charges). The mean charges are computed to be $13270.42, and the standard deviation of charges is $12105.48. There are potential outliers beyond $50000.

Figure 1: Distribution of Charges



Method 1 - Linear Regression:

First, we will build a modal for insurance charge based on linear regression. Linear regression tries to modal a relationship between dependent variables and independent variable by fitting the data into a linear line. Thus, we can use the data we have on hand (dependent variables) to predict the value we need (independent variable).

A linear regression line has an equation of the form $\mathbf{Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p}$, where Y is the dependent variable, X_1 through X_p are the independent variables we will use to predict Y, b_0 is the value when all of the independent variables are equal to zero, and b_1 through b_p are regression coefficients. In our case, Y will be **charges**, and X_1 through X_p will be age, sex, bmi, children, smoker, and regions.

Linear regression seeks to find slopes and intercept (b in the equation) which can minimize the square of the differences between predicted values and observed values. A cost function is represented below:

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2$$

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

To build a linear regression model for our case in python, we first need to clear our data. Since we are focusing on the **charges** variable (dependent variable), we need to separate this variable from the rest. We also need to handle categorical variables (sex, smoker and region) by modifying each individual category in the variables into new variables. After we do this, our variables should become age, sex_male, bmi, children, smoker_yes, region_northwest, region_southeast, and region_southwest. Then, we separate our datasets into two: one for training (80% of the data), and one for testing (20% of the data). Training data will be used to build the regression model, and testing data will be used to test our model's accuracy. Subsequently, we can apply the linear regression model and test its accuracy by using r2 score and mean squared error. A sample of this whole process to build a linear regression model is shown below.

```
#                          Linear                          Regression

#                          import                          libraries
import                   pandas                  as                  pd
importnumpyas np


#                                read                              data
df = pd.read_csv("insurance.csv")


#              separate              dependent              variable
x                      =                     df.drop("charges",axis=1)
y = df["charges"]


#               handle                categorical                data
x = pd.get_dummies(x, drop_first=True)


#     split     data     into     train     and     test     sets
fromsklearn.model_selectionimporttrain_test_split
x_train,x_test,y_train,y_test                                      =
train_test_split(x,y,test_size=0.2,random_state=42)


#          apply          linear          regression          model
```


ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

```
fromsklearn.linear_modelimportLinearRegression
LR                          =                          LinearRegression()
LR.fit(x_train,y_train)

print(LR.intercept_)

print(LR.coef_)


#                 test                regression                model
y_training_prediction               =              LR.predict(x_train)
y_testing_prediction = LR.predict(x_test)


#                import                r2_score                module
fromsklearn.metricsimport                                      r2_score
fromsklearn.metricsimportmean_squared_error


#                            accuracy                            score
score1=r2_score(y_train,y_training_prediction)
score2=r2_score(y_test,y_testing_prediction)
print("r2  score  for  training  is",score1,"r2  score  for  testing
is",score2)
print("Mean               squared               error               is
",mean_squared_error(y_test,y_testing_prediction))
```

The model we get from this process is $Y = -11931 + 257X_1 + 337X_2 + 425X_3 - 19X_4 + 23651X_5 - 371X_6 - 658X_7 - 810X_8$, or we can represent it as Charges = -11931 + 257*Age + 337*BMI + 425*Children - 19*Sex_Male + 23651*Smoker_Yes - 371*Region_Northwest - 658*Region_Southeast - 810*Region_Southwest. For categorical variables in this function (sex_male, smoker_yes, region_northwest, region_southeast, and region_southwest), x equals 1 when the variable is a true description and equals 0 when the variable is a false description.

The r2 score has a value between 0 to 1, which measures the proportion of variance in the dependent variable predicted from the independent variables. The higher the r2 score, the more accurate the model is. Our linear regression model has a r2 score of 0.74 for training data and 0.78 for testing data, which means that our model is accurate enough. Mean squared error is the average of the square between predicted values and observed values. The lower the mean

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

squared error, the more accurate the model is. Our linear regression model has a mean squared error of 33596915, which is relatively small as insurance charges are big. Hence, we can conclude that the linear regression model **Charges = -11931 + 257*Age + 337*BMI + 425*Children - 19*Sex_Male + 23651*Smoker_Yes - 371*Region_Northwest - 658*Region_Southeast - 810*Region_Southwest**is true to represent the relationship between insurance cost and the variables of interests.

Method 2 - Lasso Regression:

Lasso regression is a type of linear regression that uses shrinkage. Its cost function is shown below:

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

As illustrated in the function, lasso regression adds a penalty equivalent to the magnitude of the coefficient. When lambda increases, more coefficients are set to zero and eliminated. As lambda equals zero, the objective function will be the same as regular linear regression objective function. Since the increase in lambda causes bias and decrease in lambda causes variance, a good value for lambda is important.

To build a lasso regression model for our case, we do the same steps as linear regression. The only difference is that instead of applying linear regression model, we apply lasso regression model and try different values for lambda.

```
#          apply          lasso          regression          model
fromsklearn.linear_modelimportLinearRegression
LR                        =                        Lasso(alpha=90)
LR.fit(x_train,y_train)
print(LR.intercept_)
print(LR.coef_)
```

After trying different values for alpha, we find that 90 fits our data well. Hence, our model becomes $Y = -11833 + 256X_1 + 325X_2 + 369X_3 - 0X_4 + 23102X_5 - 0X_6 - 0X_7 - 0X_8$, or represented as Charges = -11833 + 256*Age + 325*BMI + 369*Children + 23102*Smoker_Yes.

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

Here, we eliminate variables sex and regions. As illustrated in Figure 2 and Figure 3, people of different sex and from different regions do not have a large variance in insurance cost, so the elimination by lasso regression is reasonable.

Figure 2: Bar Chart of Sex and Charges

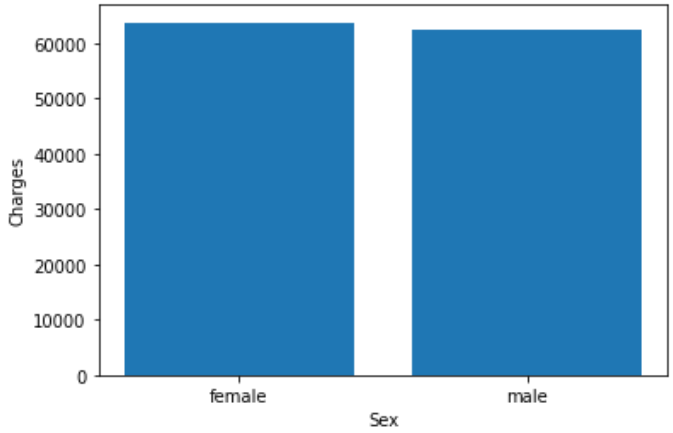
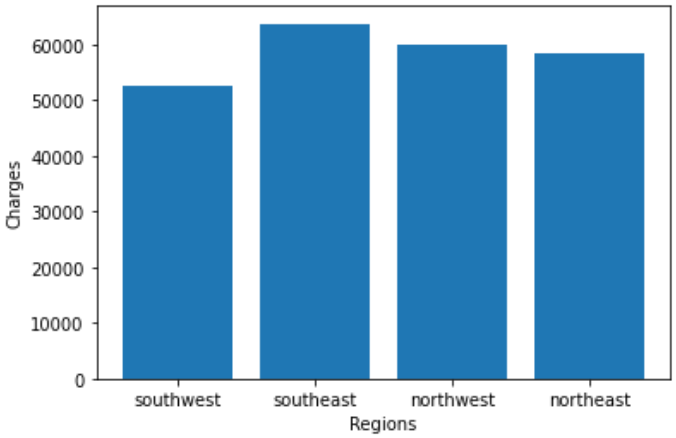


Figure 3: Bar Chart of Regions and Charges



Our lasso regression model has a r2 score of 0.74 for training data and 0.78 for testing data. Mean squared error for this model is 34212931. These measures show that the model is accurate.

Method 3 - Ridge Regression:

Ridge regression is also a type of linear regression. It is similar to lasso regression as it takes account of the penalty equivalent. However, instead of taking the magnitude of coefficient, ridge

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

regression uses the square of magnitude of coefficients. Below is the cost function of ridge regression:

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

To build a ridge regression model, we also use the same process as the linear regression, illustrated in Method 1. However, we modify the applying linear regression model step into applying ridge regression model.

```
# apply ridge regression model

fromsklearn.linear_modelimport                                    Ridge
LR                          =                                Ridge(alpha=2)
LR.fit(x_train,y_train)
print(LR.intercept_)
print(LR.coef_)
```

Here, we set alpha equal to 2 and we get the model Y = - 11869 + 257X_1 + 337X_2 + 426X_3 - 3X_4 + 23381X_5 - 362X_6- 631X_7 - 798X_8, or represent it as Charges = -11869 + 257*Age + 337*BMI + 426*Children - 3*Sex_Male + 23381*Smoker_Yes - 362*Region_Northwest - 631*Region_Southeast - 798*Region_Southwest. When we take account of the penalty equivalent using the square of magnitude of coefficients, no variables can be eliminated. However, we observe that sex only has a coeffient of -3. Compared to other variables, sex is less relavant to insurance expenses.

The r2 score we get for training data using ridge regression is 0.74 and for testing data is 0.78. The mean squared error for this model is 33698329. These all shows that our ridge regression model is true to represent the relationship between charges and age, bmi, children, sex, smoker, and region.

Comparison:

Linear regression represents the relationship between dependent variables and independent variable. It regards all dependent variables at equivalent importance, and because of that, we get

ISSN: 2455-8834

Volume:07, Issue:02 "February 2022"

a function with large coefficients in some variables. Ridge regression takes the penalty equivalent into account when building the model. It means that if the coefficient is large, the function will be penalized. Hence, it shrinks the coefficient of dependent variables and reduces the complexity of the model. Compared to ridge regression, lasso regression takes only the magnitude of coefficients rather than the square of the magnitude of coefficients to penalize the function. In this case, it not only regularizes the coefficients, but also eliminates some of the less relevant variabels.

Conclusion:

In our research, we try to establish a relationship between several features (age, bmi, sex, children, smoker, and region) and insurance charges. For simplicity and accuracy reasons, the model offered by lasso regression is the best. (**Charges = -11833 + 256*Age + 325*BMI + 369*Children + 23102*Smoker_Yes**) As offered by this function, we can better understand the insurance cost as we can see which variable is more important to insurance cost (large coefficient), and which variable is not. We, as insurance seekers, can then apply our situations to the model to examine what our insurance expense will be. Insurance companies can also use this model to determine the cost of insurance for individuals.

Reference

[1] Seber, G.A. and Lee, A.J., 2012. *Linear regression analysis*. John Wiley & Sons.

[2] Ranstam, J. and Cook, J.A., 2018. LASSO regression. *Journal of British Surgery*, *105*(10), pp.1348-1348.

[3] Marquardt, D.W. and Snee, R.D., 1975. Ridge regression in practice. *The American Statistician*, *29*(1), pp.3-20.

[4] Kira, K. and Rendell, L.A., 1992. A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249-256). Morgan Kaufmann.

[5] Hackeling, G., 2017. *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd.

[6]Akossou, A.Y.J. and Palm, R., 2013. Impact of data structure on the estimators R-square and adjusted R-square in linear regression. *Int. J. Math. Comput*, *20*(3), pp.84-93.

