CIS – 490: Machine Learning

Learning Activity 2

Name: Pranav Vinod

UMassD ID: 01984464

PART 1

Q1. Describe Loss Function for linear regression, SSL/SSR/RSS.

Ans. The Loss function or Residual Sum of Squares (RSS) or Sum of Squared Residue (SSR) or Sum of Squared Loss (SSL) is the function that determines the difference between the observed value of the outcome and the predicted value of the outcome. It is a measure of the cost of predicting the outcome using a model, in this case, the linear model.

Coefficients of the regression line are found such that the loss function is minimized.

Let , be the predicted outcome,

Where and are the slope and intercept respectively and is the error term.

And y be the observed outcome

is the residual (i.e., loss) in the prediction of outcome

Then,

SSR/RSS/SSL =

=

Q2. Describe the least squares estimation for linear regression.

Ans. Having defined the Loss Function (SSR) as

SSR =

We need to find the coefficients and such that the loss function is minimized, and the best fit line can be achieved.

To do so, we can differentiate SSR with respect to and and equate it to 0.

After solving we get,

=

=

Q3. Discuss Accuracy Checking/Model Fit numeric criterion for linear regression.

Ans. Now that we have calculated a model to predict our outcome we need to check how well the model fits with our observed values or how correct the model is.

To do so, we can compute certain numeric quantities to check the accuracy of the model calculated.

There are 4 main numeric quantities to compute, these are:

1. Residual Standard Error (RSE)
2. F – Statistic
3. Mean Squared Error (MSE)

Residual Standard Error (RSE)

Having defined RSS as

We now define

RSE =

Where n is the number of observations,

And p is the number of parameters,

For a simple linear regression model,

RSE =

RSE is a measure of lack of fit of the model to the data.

A small value of RSE indicates that the model fits the data well and a large value of RSE indicates that the predicted value from the model is far away from the observed value of the outcome variable.

Statistic

The statistic is defined as

= 1 – RSS/TSS

Where RSS =

And TSS = is the total sum of squares.

Here RSS measures the amount of variability in *y* that cannot be explained after performing regression.

Whereas TSS measures the total amount of variability in *y* which is inherent in the outcome before the regression is performed.

So measures the proportion of variability in *y* that can be explained by *x*. Since it is a proportion, it always lies between 0 and 1.

A value close to 1 indicates that a large amount of variability in the outcome can be explained by the predictor variable. Whereas a value close to 0 means that the model does not explain the variability in *y* well.

F – Statistic

The F – Statistic is defined as

F =

Here TSS and RSS mean the usual quantities and p and n are the number of parameters and observations respectively.

A value of F larger than 1 indicates that at least one of the predictor variables is related to the outcome.

Although if n is large enough, a F value slightly larger than 1 might be enough to indicate a strong relationship between one of the predictors and the outcome.

For a small value of n, a significantly large value of F would be needed to make the assertion.

Mean Squared Error (MSE)

MSE is defined as

MSE = RSS/n

It is a generic criterion for any regression model.

Q4. Describe training set, testing set, training error and testing error.

Ans. To run a regression model using a machine, we need to divide our dataset into 2 parts:

1. Training set and Training Error

This part of the dataset is used to construct the model by using the known values of the predictor and outcome to find out appropriate coefficients that give the best fit regression line.

Training error is the generic MSE, RSE, F-Statistic and calculated using the model created using training set.

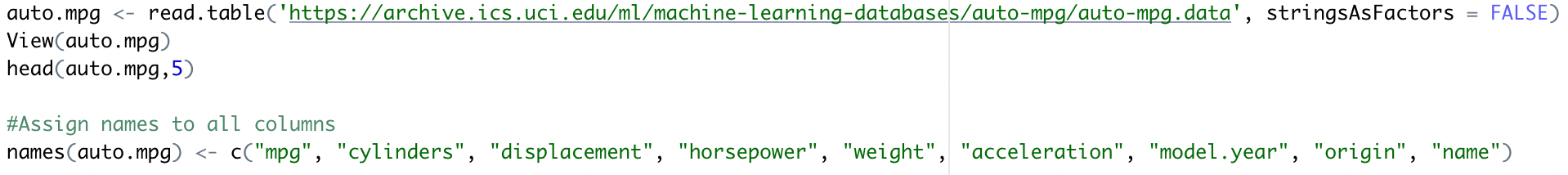
1. Testing set and Testing error

This part of the dataset is used to test our created model from the training set against known values of the outcome.

The testing error is the error in the predicted outcome from the model as compared to the known values of the outcome in the testing data. We use the generic MSE to compute this error.

PART 2

Q1. Download Auto MPG data from  <https://archive.ics.uci.edu/ml/datasets/auto+mpg;refer> to Lecture 6 slides and posted instruction file, called ”R\_LinearRegression\_LS6” at mycourses for running linear regression in R:

Ans. 

Q2. Run simple linear regression on training, test your trained model on the testing data, and entire data (50% for training; 50% for testing; for rng seed, use ”490” with default random generator "Mersenne-Twister to replicate your work), draw the best fit line on a scatter plot; output all model fit statistics.

Ans.

Graphical user interface, text, application

Description automatically generated

Running the model on training set

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Now we run the trained model on testing set,

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Plotting the best fit line,

A picture containing text

Description automatically generated

Chart, scatter chart

Description automatically generated

Now we run simple regression on the entire data,

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

Q3. Run multiple linear regression on training set (50%), test your trained model on the testing data (50%), and entire data; output all model fit statistics.

Ans. Running multiple linear regression on training set,

Text

Description automatically generated with medium confidence

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Now we run the trained multiple regression model on testing dataset,

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Now plotting against each variable

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Now running the multiple regression model on the entire dataset,

Text

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Q4. Run the parsimonious multiple linear regression model on training set (50%), test your trained model on the testing data (50%), and entire data. output all model fit statistics. Refer to Slide 29 in LS6, report a summary table of all model fit statistics generated from a, b and c, and write your final parsimonious model based on the entire dataset (e.g., Y = 0.5 + .02X).

Ans. From the multiple linear regression model, we can see that there are only two predictors who have significant impact on the outcome.

These are the weight and model year.

So now, we run the multiple linear regression model using only those two variables.

Running the model on training set,

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

Now running the model on testing data,

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Plotting,

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Now running the parsimonious model on the entire dataset,

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Summary table comparing statistics between simple and multiple regression for training, testing, and full dataset (omitting RMSE from the table):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | Simple Linear Regression | | | Multiple Regression | | | Multiple Regression  (Only significant variables) | | |
| Train | Test | All | Train | Test | All | Train | Test | All |
| MSE | 18.6 | 18.9 | 18.8 | 10.5 | 13.07 | 11.6 | 10.5 | 12.9 | 11.7 |
| RSS | 3707.7 | 3775.8 | 7474.8 | 2084.5 | 2601.0 | 4639.8 | 2099.1 | 2582.8 | 4659.8 |
| RSE | 4.33 | - | 4.34 | 3.28 | - | 3.44 | 3.27 | - | 3.43 |
| R^2 | 0.684 | - | 0.691 | 0.8222 | - | 0.808 | 0.821 | - | 0.808 |
| Adj. R^2 | 0.682 | - | 0.691 | 0.818 | - | 0.806 | 0.819 | - | 0.807 |
| F-Statistic | 426.8 | - | 888.9 | 178.8 | - | 331.4 | 450.1 | - | 830.4 |

The final parsimonious model based on the entire dataset is:

We can see that there is a negative relation between MPG and weight, which means that as the weight increases the miles per gallon of a vehicle decreases. This makes sense as a heavier vehicle would consume more fuel to run for a mile than a lighter vehicle.

Also, we can see that as the model year of a vehicle increases, it’s miles per gallon also increases. This is true as a newer vehicle would run for more miles per gallon then an older vehicle.