Introduction to Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occurin almost every field, including engineering, the physical and chemical sciences,economics, management, life and biological sciences, and the social sciences. In fact,regression analysis may be the most widely used statistical technique.

Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

y = β0 + β1x

where β0 and β1 represent two model parameters (β is the Greek letter beta). (This use of β has nothing to do with the β we used to describe the probability of a Type II error.) These parameters are estimated using data, and we write their point estimates as b0 and b1. When we use x to predict y, we usually call x the explanatory or predictor variable, and we call y the response.

There are 4-5 types of regression algorithms

```
1.Linear Regression
2.Polynomial Regression
3.Ridge regression
4.Lasso regression
5.ElasticNet
```

What is Independent and Dependent Variables?

Independent variables (also referred to as Features) are the input for a process that is being analyzes. Dependent variables are the output of the process.

Dependent variables are nothing but the variable which holds the phenomena which we are studying. Independent variables are the ones which through we are trying to explain the value or effect of the output variable (dependent variable) by creating a relationship between an independent and dependent variable.

What is splitting the data into train and test??

It is the splitting of a dataset into multiple parts. We train our model using one part and test its effectiveness on another.

In practice, data usually will be split randomly 70-30 or 80-20 into train and test datasets respectively in statistical modeling, in which training data utilized for building the model and its effectiveness will be checked on test data.

In the following code, we split the original data into train and test data by 80 percent – 20 percent.

Data splitting is the process of splitting data into 3 sets:

```
Data which we use to design our models (Training set)
Data which we use to refine our models (Validation set)
Data which we use to test our models (Testing set)
```

If we do not split our data, we might test our model with the same data that we use to train our model.

What is a Training Set?

The training set is the set of data we analyse (train on) to design the rules in the model.

A training set is also known as the in-sample data or training data.

What is a Validation Set?

The validation set is a set of data that we did not use when training our model that we use to assess how well these rules perform on new data.

It is also a set we use to tune parameters and input features for our model so that it gives us what we think is the best performance possible for new data.

What is a Test Set?

The test set is a set of data we did not use to train our model or use in the validation set to inform our choice of parameters/input features.

We will use it as a final test once we have decided on our final model, to get the best possible estimate of how successful our model will be when used on entirely new data.

A test set is also known as the out-of-sample data or test data.

Why do we need to split our data?

To prevent look-ahead bias, overfitting and underfitting.

1. Look-ahead bias: Building a model based on data that is not supposed to be known.
2. Overfitting: This is the process of designing a model that adapts so closely to historical data that it becomes ineffective in the future.
3. Underfitting: This is the process of designing a model that adapts so loosely to historical data that it becomes ineffective in the future.

What is evaluation of regression models?

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at MSE, MAE, R-squared, Adjusted R-squared, and RMSE. The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value.

1.Mean Squared Error

MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.

2.Mean Absolute Error

We know that an error basically is the absolute difference between the actual or true values and the values that are predicted. Absolute difference means that if the result has a negative sign, it is ignored.

Hence, MAE = True values – Predicted values

MAE takes the average of this error from every sample in a dataset and gives the output

3.Root Mean Squared Error

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

4.R Squared

It is also known as the coefficient of determination. This metric gives an indication of how good a model fits a given dataset. It indicates how close the regression line (i.e the predicted values plotted) is to the actual data values. The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

In [ ]:

1. Linear Regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Assumption in linear Regression

1.Linear Relation

There exists a linear relationship between the independent variable, x, and the dependent variable, y.

Methods make use when there is no linear Relation

```
1.Apply some non linear transformation -> log ,square root
2.To add another independant var , x=> x^2
```

2.Homoscedasticity

Residuals should have the constant variance at every level of x This is known as homoscedasticity. When this is not the case, the residuals are said to suffer from heteroscedasticity.

There are three common ways to fix heteroscedasticity:

```
1.Transform the dependent variable.
2.Redefine the dependent variable.
3.Use weighted regression.
```

plot the fitted values vs residuals

3. Normality of Residuals

4.Mean of Residuals

Residuals as we know are the differences between the true value and the predicted value. One of the assumptions of linear regression is that the mean of the residuals should be zero.

5.Multicolinearity

In regression, multicollinearity refers to the extent to which independent variables are correlated. Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don't need to understand the role of each independent variable, you don't need to reduce severe multicollinearity.

In [ ]:

2.Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x). Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(y | x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression

In [ ]:

3.Ridge Regression

Tikhonov regularization, named for Andrey Tikhonov, is a method of regularization of ill-posed problems. Ridge regression is a special case of Tikhonov regularization in which all parameters are regularized equally. Ridge regression is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias (see bias–variance tradeoff).

In [ ]:

4.Lasso Regression

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. It was originally introduced in geophysics, and later by Robert Tibshirani, who coined the term.

In [ ]:

5.ElasticNet

Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions. The coefficients of the model are found via an optimization process that seeks to minimize the sum squared error between the predictions (yhat) and the expected target values (y). A problem with linear regression is that estimated coefficients of the model can become large, making the model sensitive to inputs and possibly unstable. This is particularly true for problems with few observations (samples) or more samples (n) than input predictors (p) or variables (so-called p >> n problems).

One approach to addressing the stability of regression models is to change the loss function to include additional costs for a model that has large coefficients. Linear regression models that use these modified loss functions during training are referred to collectively as penalized linear regression.

ElasticNet CV

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set X_test, y_test. Note that the word "experiment" is not intended to denote academic use only, because even in commercial settings machine learning usually starts out experimentally. Here is a flowchart of typical cross validation workflow in model training.

In [ ]: