Linear Regression

Linear Regression

Linear regression is probably the simplest approach for statistical learning. It is a good starting point for more advanced approaches, and in fact, many fancy statistical learning techniques can be seen as an extension of linear regression. Therefore, understanding this simple model will build a good base before moving on to more complex approaches.

Building Model from Data

- There are different ways to understand and quantify relationship between variables, e.g., is there a relationship between age and cholesterol levels? Is there any relation between central bank meeting and stock market?
- Formal ways to describe, encode and test if and how one or more variables relate to others is to build and evaluate models from the data. These models describe important relationship in the data, including the strength and direction – positive or negative of the relation.
- Models quantify the relationship between variables. The models can represent linear and nonlinear relationships in the data.
- Linear regression is one of easiest algorithms in machine learning

Building Model from Data

- They can also be used to confirm hypothesis about relationships.
- All these uses help to summarize and understand the data.
- However, one of the most widely used applications of a model is to make predictions.
- A model is usually built to predict values for a specific variable, e.g., were a data set composed of historical data containing attributes of pharmaceuticals and their observed side effects to be collected, a model can be generated from this data to predict the side effects from the pharmaceutical’s attributes.

Building Model from Data

Linear regression is a parametric technique used to predict the value of a continuous variable output Y based on one or more input predictor variables X.

It is parametric in nature because it makes certain assumptions based on the data set.

If the data set follows those assumptions, regression gives very good results. Otherwise, it struggles to provide convincing accuracy.

The aim is to establish a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, we can use this formula to estimate the value of the response Y, when only the predictors (Xs) values are known.

Building Model from Data

A variable that a model predicts is often referred to as **response variable** or **dependent-variable**. The variable encoded in the model used in predicting this response are referred to as the **independent variables** or **x-variables**.

.

Linear Regression Model

- It is a technique which explains the degree of relationship between two or more variables using a best fit line/plane.

- Linear regression generates a linear model to describe a relationship between one or more independent variables (x_1, …, x_p) and response variable y.

- In regression, the response variable must be a continuous variable

- The predictors (x_1, …, x_p) can be continuous, discrete or categorical variables.

Building Model from Data

- Although response variable is known, when building models it is not always apparent beforehand which variable should be used as independent variables. Therefore selection of the independent variables is an important step in building the model.

- Knowledge of the problem can guide the choice of variables to use in models. Alternatively, we could build multiple models with different combinations of independent variables and select the best fitting model.

- Another issue to consider when selecting independent variable is the relationship between the independent variables.

Building Model from Data

- Combinations of variables that have strong relationships to each other should be avoided since they will be essentially encoding the same relationship to the response.

- Including all the variables from each group of strongly related variables produces overly complex models and can produce results that are difficult to interpret.

- In developing a model, it may also be necessary to use derived variables, that is , a new variable that is a function of one or more variables. For example, because most modeling methods require numeric data, if a data set has nominal variable that will be used in the model, the values of these variables must be transformed into numbers.

Linear Regression

Linear regression is very good to answer the following questions:

- Is there a relationship between 2 variables?
- How strong is the relationship?
- Which variable contributes the most?
- How accurately can we estimate the effect of each variable?
- How accurately can we predict the target?
- Is the relationship linear?

Assumptions of Linear Regression Model

1. **Linear Relationship** : Linear regression needs a linear relationship between the dependent (DV) and independent variables (IV). By linear, it means that the change in dependent variable by 1 unit change in independent variable is constant.

2. **Normality of Residual** : Linear regression requires residuals should be normally distributed.

3. **No Outlier Problem**

4. **Multicollinearity** : It means there is a high correlation between independent variables. The linear regression model MUST NOT be faced with problem of multicollinearity. If variables are correlated, it becomes extremely difficult for the model to determine the true effects of IVs on DV

Assumptions of Linear Regression Model

5. Independence of error terms - No Autocorrelation

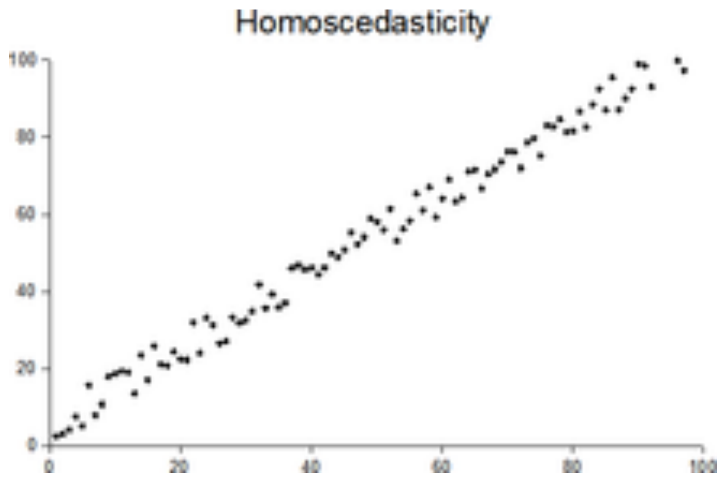
It states that the errors associated with one observation are not correlated with the errors of any other observation. It is a problem when you use time series data. It drastically affects the regression coefficients and standard error values since they are based on the assumptions of uncorrelated error terms. Suppose you have collected data from labors in eight different districts. It is likely that the labors within each district will tend to be more like one another that labors from different districts, that is, their errors are not independent.

Presence of these assumptions make regression quite restrictive. By restrictive means that the performance of a regression model is conditioned on fulfillment of these assumptions.

Assumptions of Linear Regression Model

6. Homoscedasticity – Constant Variance of residuals

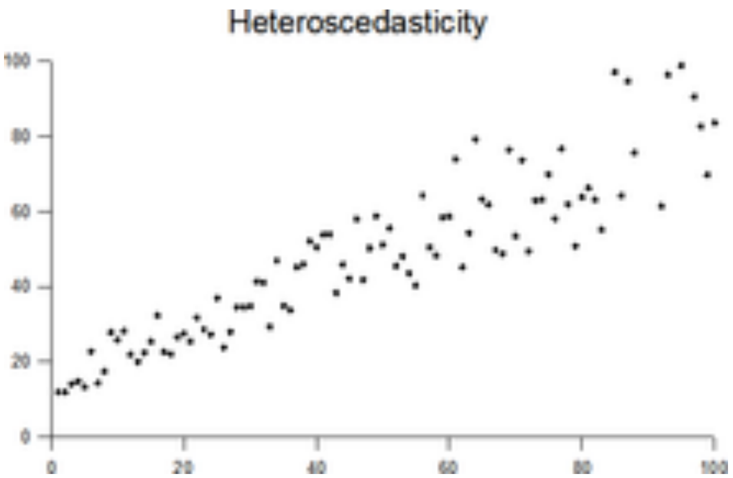
In English: Residuals evenly distributed around the mean or residuals are approximately equal for all predicted dependent variable values.



Residuals:

– The difference in predicted value of the dependent variable and the observed value of the dependent variable

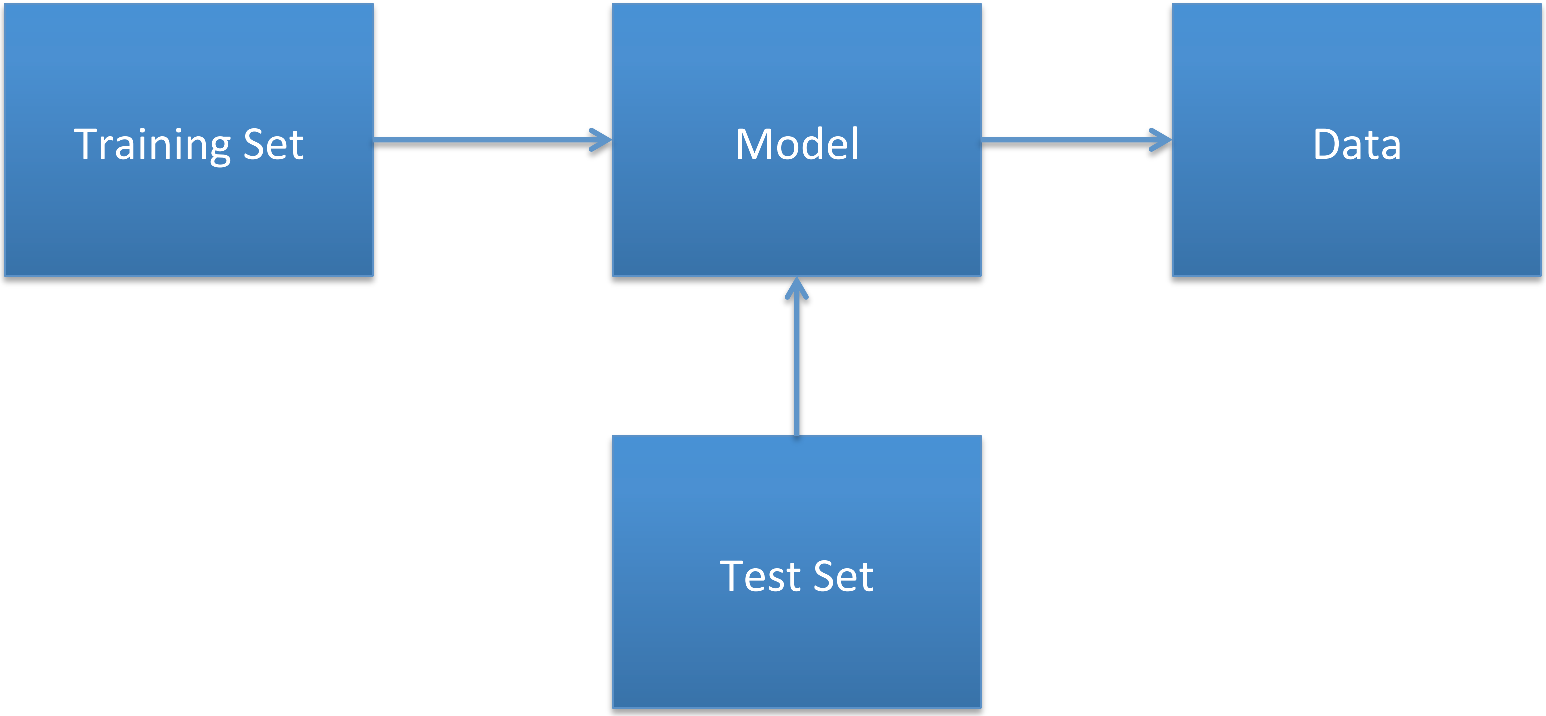
– Y_predicted – Y_actual



Building Model from Data

- The data set used to build a model is referred to as the training set. To test the performance of a generated model, a test set with observations different from those in the training set is used to test how well the model performs. The model uses the values of each observation in the test set to predict a value for the response variable.

Building Model from Data



Linear Regression Model

Strengths:

- By far the most common approach for modeling numeric data
- Can be adapted to model almost any data
- Provides estimates of the strength and size of the relationships among the feature and the outcome

Weaknesses:

- Makes strong assumptions about the data
- The model’s form must be specified by the user in advance
- Does not do well with missing data
- Only works with numeric features, so categorical data require extra processing
- Requires some knowledge of statistics to understand the model.

Linear Regression Model

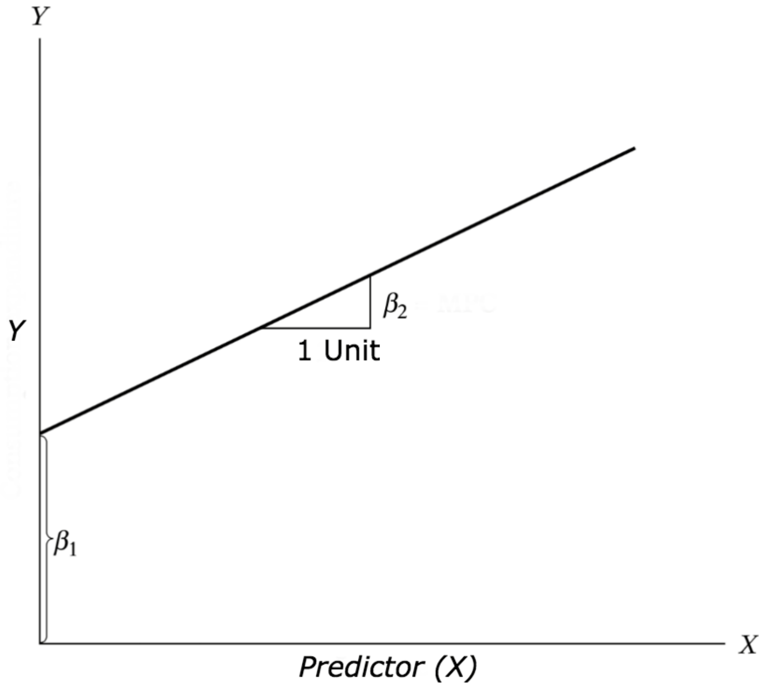
- A simple linear regression is the most basic model. It just involves two variables and is modeled as a linear relationship with an error term

$$Y = \beta_1 + \beta_2 X + \varepsilon = h(X) + \varepsilon \quad \Rightarrow \varepsilon = Y - h(X)$$

- Where β_1 is the intercept and β_2 is the slope. Collectively, they are called regression coefficients.
- ε is the random error term, the part of Y the regression model is unable to explain. For ideal model, this should be random and should not be dependent on any input. No matter how powerful the algorithm we choose, there will always remain an ε irreducible error which reminds us that the future is uncertain.

Linear Regression Model

- β_1 is the expected mean value of dependent variable when all independent variables are equal to 0.
- β_2 slope represents the amount by which dependent variable changes if we change X by one unit.



Linear Regression Model

- These coefficients are what we need in order to make predictions from our model.

- We are given the data x and y , our mission is to fit the model which will give us the best estimates for *betas*. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error are as small as possible. Error is the distance between the point to the regression line.

- To do this, regression uses a technique know as Ordinary Least Square (OLS)

Linear Regression Model

OLS technique tries to reduce the sum of squared errors Σ(actua;(y)-predicted(y))^2 by finding the best possible value of regression coefficients

why OLS? Let's see.

- It uses squared error which has nice mathematical properties, thereby making it easier to differentiate and compute gradient descent.
- OLS is easy to analyze and computationally faster, i.e. it can be quickly applied to data sets having 1000s of features.
- Interpretation of OLS is much easier than other regression techniques.

Linear Regression Model

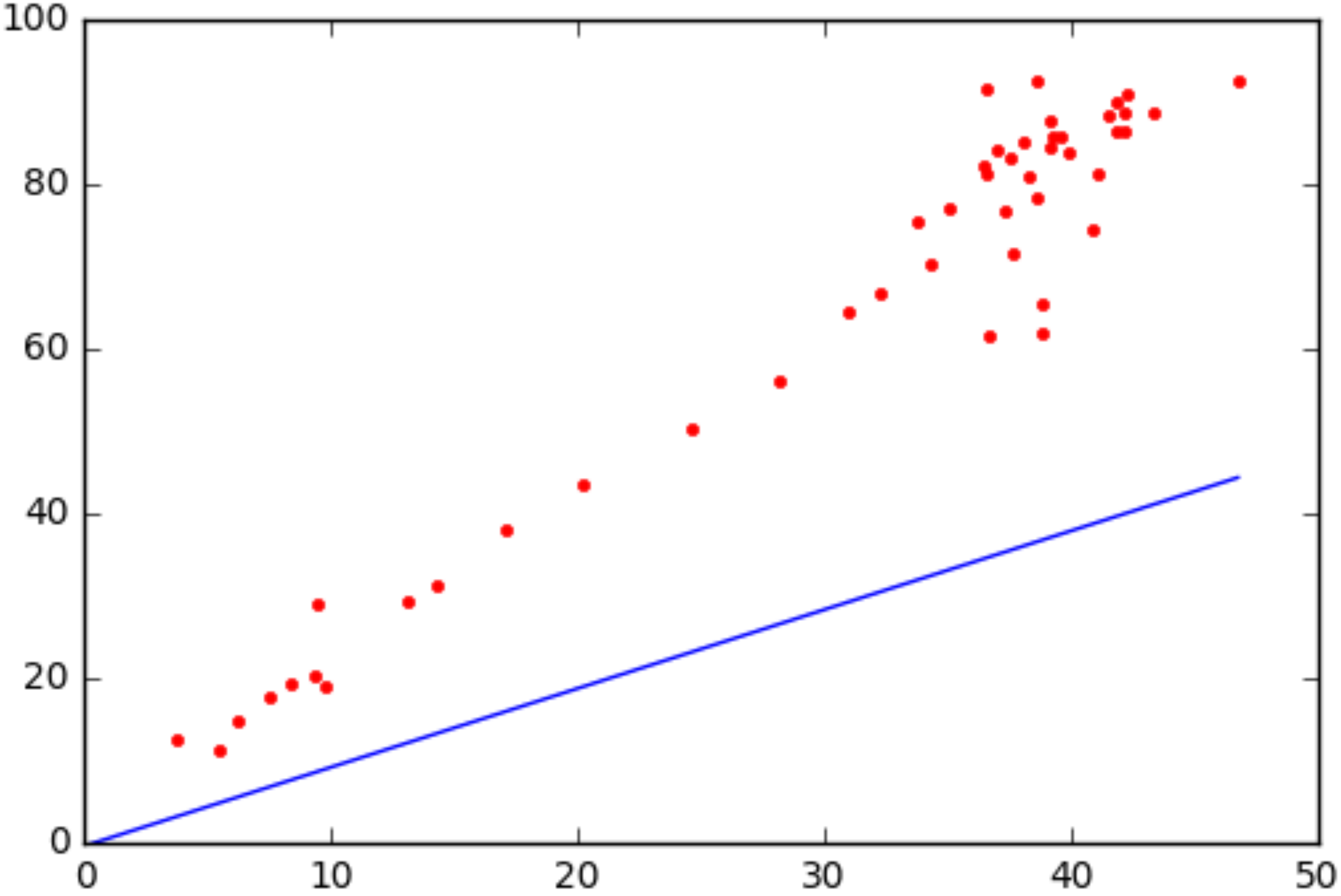
The error is easily calculated as:

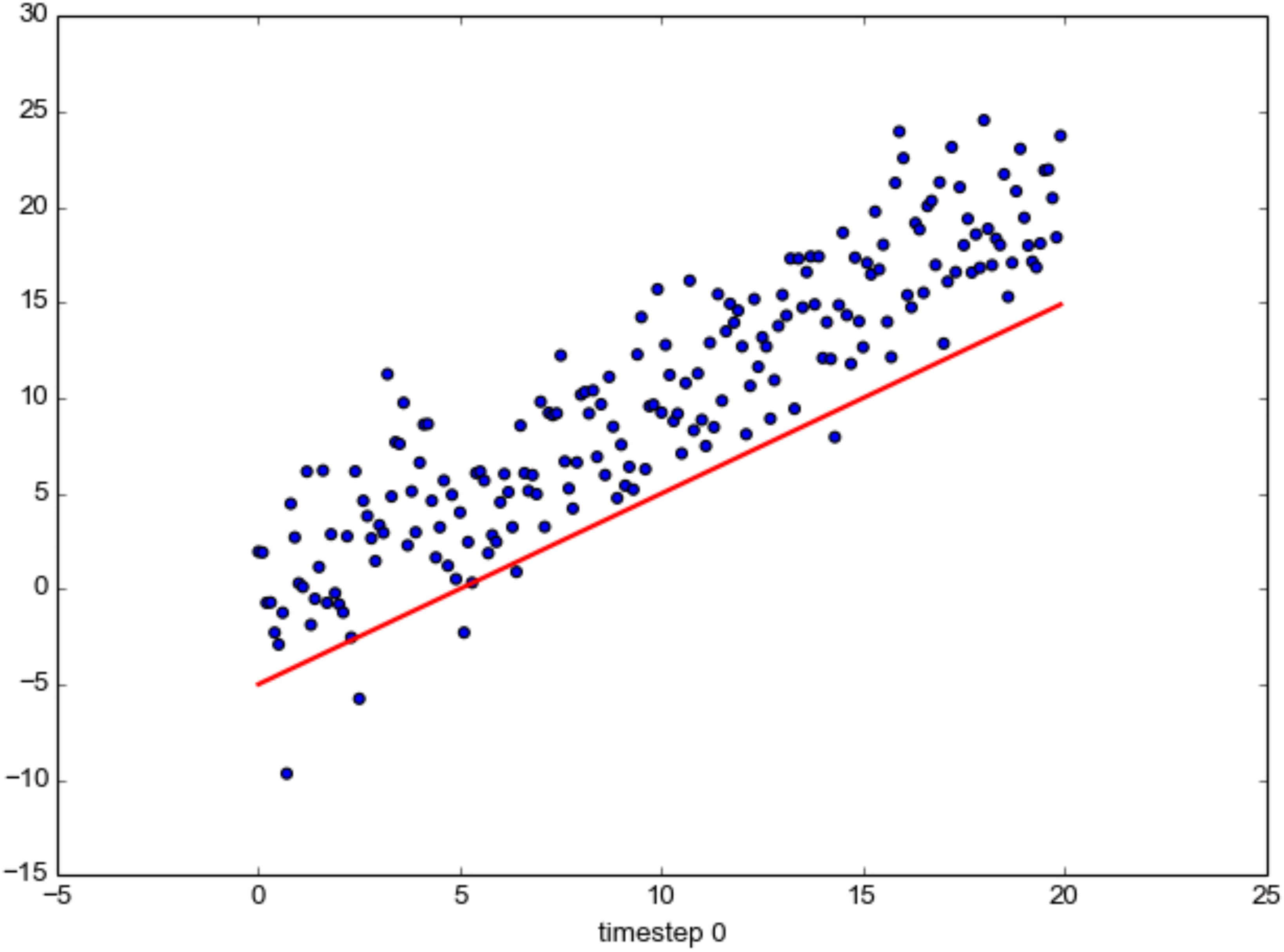
$$e_i = y_i - \hat{y}_i$$

We define the squared error or cost function, J as:

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

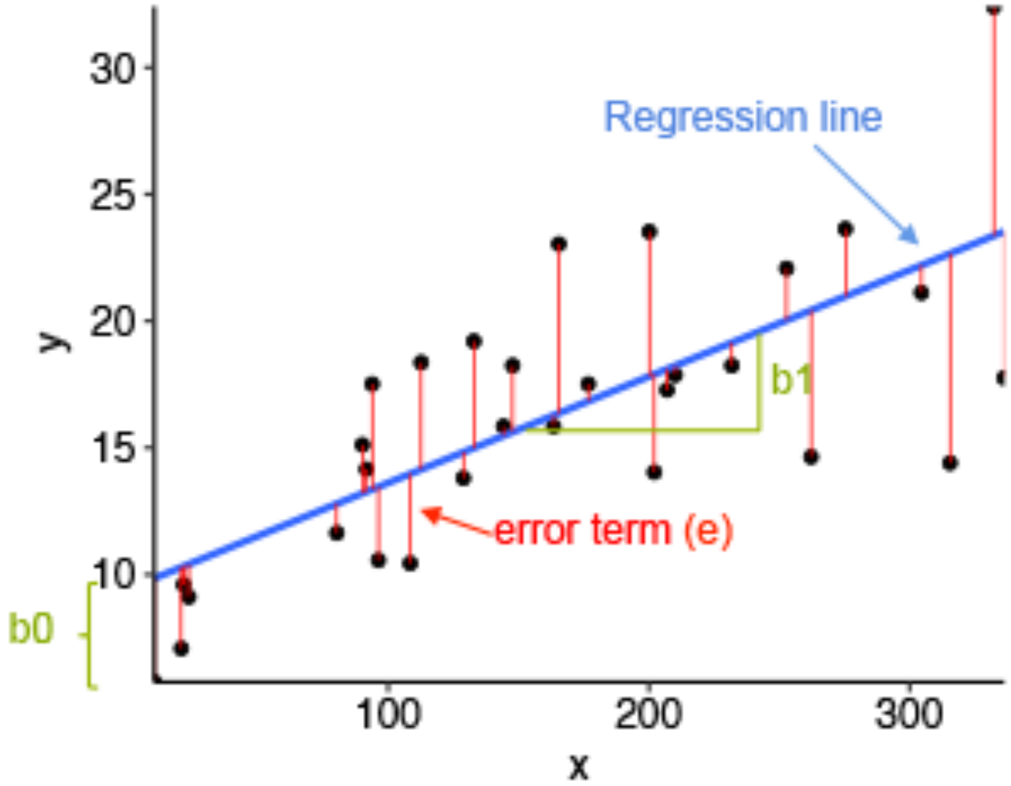
And our task to find the values of β_0 and β_1 for which $J(\beta)$ is minimum. To find the parameters, we need to minimize the least squares or the sum of squared errors. This is called a learning procedure. We can find these using different approaches. One is called Ordinary Least Square Method and other one called Gradient Descent Approach.


100
80
60
40
20
0
0
10
20
30
40
50


30
25
20
15
10
5
0
−5
−10
−15
−5
0
5
10
15
20
25
timestep 0

Linear Regression Model

Linear regression is base on least square estimation which says that regression coefficients should be chosen in such a way that it minimizes the sum of squared distances of each observed response to its fitted value.



Linear Regression Model

What the linear regression technique does is , it finds the best possible line which fits that training set and then predicts for any unseen data.

Linear regression finds the parameters of that line which best fits the data, i.e., slope (β_1) and intercept (β_0).

Linear Regression Model

Without going into the mathematical details, one can finally estimate the coefficients with the following equations

$$\beta_2 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\beta_1 = \overline{y} - \beta_2 \overline{x}$$

Where x bar and y bar represent the mean

Linear Regression Model

Now, determine estimate of β's using least square method and the solution is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

This is also know as normal equation solution.

Knowing the estimates of β's, the multiple regression model can be estimated as

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

Linear Regression Model

In real life situations, there ill never be a single feature to predict a target. So, do we perform linear regression on one feature at a time? Of course not. We simply perform multiple linear regression. Clearly, it is nothing but an extension of Simple linear regression.

The equation is very similar to simple linear regression; simply add the number of predictors and their corresponding coefficients:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

Linear Regression Model

- Prior to any modeling, the data should always be inspected for:
  - Data entry errors
  - Missing values
  - Outliers
  - Unusual (e.g., asymmetric) distribution
  - Non-linear bivariate relationships
  - Unexpected patterns

Linear Regression Model

- We can resort to:
  - Numerical summaries
  - Correlations
  - etc.
- Graphical summaries:
  - boxplots
  - Histograms
  - Scatterplots
  - etc.

Linear Regressions – The things you can do

Build the model

- SimpleLinearRegression
- MultipleLinearRegression

Variations

- Choosing the ‘best’ regression variables
- Regression on a subset of the observations

Understand the output

- Interrogating the linear model statistics
- Understanding the Regression Summary

Diagnose the fit

- Plotting Regression Residuals
- Diagnosing the lm output.Goodfit?

Use the model

- Predict New Values

Estimating the relevancy of the coefficients

Now that we have coefficients, how can you tell if they are relevant to predict the target?

The best way is to find the *p-value*. The *p-value* is used to quantify statistical significance; it allows to tell whether the null hypothesis is to be rejected or not.

The null hypothesis?

For any modeling task, the hypothesis is that there is some correlation between the features and the target. The null hypothesis is therefore the opposite: There is no correlation between the features and the target. In other words, coefficients associated with the variable is equal to zero.

Estimating the relevancy of the coefficients

So, finding the *p-value* for each coefficient will tell if the variable is statistically significant to predict the target. As a general rule of thumb, if the *p-value* is less than 0.05: there is a strong relationship between the variable and the target.

Once we found that our variables are statistically significant by finding its *p-value*

t-value:

A larger t-value indicate that it is less likely that the coefficient is not equal to zero by chance. So, higher the t-value, the better.

Assess the accuracy of the model

Now, do we know if our linear model is any good

RMSE/MSE/MAE – Error metric is the crucial evaluation number. Since all these are errors, lower the number, better the model.

- **MSE** - This is mean squared error. It tends to amplify the impact of outliers on the model's accuracy.

$$MSE = \frac{(y - \hat{y})^2}{N}$$

- **MAE –** This is mean absolute error. It is robust against the effect of outliers.

$$MAE = \frac{\left| y - \hat{y} \right|}{N}$$

Assess the accuracy of the model

RMSE:

To evaluate the performance of regression model, we should look at the root mean square error. . It explains how close the actual data points are to the model’s predicted values. The lower the value, the better the model fits the data (in this case, the closer the data is to a linear relationship). It nullifies the effect of MSE by square root and provides the result in original units as data.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Assess the accuracy of the model

R Square (Coefficient of Determination)

Other metric to evaluate the performance of linear regression is R-square and most common metric to judge the performance of regression models. R^2 measures " How much change in output variable is explained by the change in the input variable. It measures the proportion of the variation in the dependent variable explained by all of independent variables in the model. Therefore, assuming linear relationship, if feature X can explain (predict) the target, then the proportion is high and the R^2 value will be close to 1. If the opposite is true, the R^2 value is then closer to 0.

$$r^2 = 1 - \frac{\text{SS Error}}{\text{SS Total}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Performance evaluation metrics

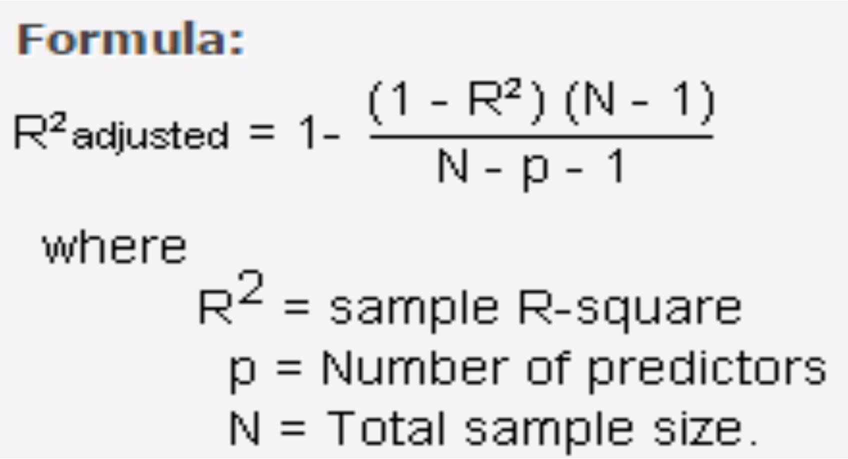
R-squared is always between 0 and 1:

- 0 indicates that the model explains NIL variability in the response data around its mean.
- 1 indicates that the model explains full variability in the response data around its mean.

In general, higher the R^2, more robust will be the model.
One disadvantage of R-squared is that it can only increase as predictor are added to the regression model. This increase is artificial when predictors are not actually improving the model's fit. To address this, we use "Adjusted R-squared"

Performance evaluation metrics

Adjusted R-square measures the proportion of variation explained by only those independent variables that really affect the dependent variable. It penalizes for adding independent variable that do not affect the dependent variable. It increases only when independent variable is significant and affects dependent variables.



Formula:

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

Assess the relevancy of a predictor

F- Statistics

It evaluates the overall significance of the model. It is the ratio of explained variance by the model by unexplained variance. It compares the full model with an intercept only (no predictors) model. Its value can range between zero and any arbitrary large number. Naturally, higher the F-statistics, better the model.

$$F = \frac{\frac{TSS - RSS}{p}}{\frac{RSS}{n-p-1}}$$

n is the number of data points and p is the number of predictors.

Assess the relevancy of a predictor

The F-statistics is calculated for the overall model, whereas the *p-value* is specific to each predictor. If there is a strong relationship, the F will be much larger that 1. Otherwise, it will be approximately equal to 1.

How large than 1 is large enough?

This is hard to answer. Usually, if there are large number of data points, F could be slightly larger than 1 and suggest a strong relationship. For small data sets, then the F value must be way larger than 1 to suggest a string relationship.

Why can’t we use the p-value in this case?

Assess the relevancy of a predictor

Why can’t we use the p-value in this case?

Since we are fitting many predictors, we need to consider a case where there are a lot of features (p is large). With a very large amount of predictors, there will always be about 5% of them that will have, by chance, a very small p-value even though they are not statistically significant. Therefore, we use the F-statistic to avoid considering unimportant predictors as significant predictors.

Linear Regression Model

- .Once we have built the model, we need to check the following things
  - Is the model statistically significant? (F statistics)
  - Are the coefficients significant (t statistics and p-values)
  - Is the model useful (R^2 value)
  - Does the model fit the data well? (plot residuals)
  - Does the data satisfy the assumptions?

How can you improve the accuracy of a model?

There is little you can do when your data violates assumptions made in the regression model. An obvious solution is to use another algorithms which capture non-linearity quite well. But if you are adamant at using regression, following are some tips you can implement:

1. If your data is suffering from non-linearity, transform the IVs using sqrt, log, square, etc.

2. If your data is suffering from heteroskedasticity, transform the DV using sqrt, log, square, etc. Also, you can use weighted least square method to tackle this problem.

How can you improve the accuracy of a model?

3. If your data is suffering from multicollinearity, use a correlation matrix to check correlated variables. Let's say variables A and B are highly correlated. Now, instead of removing one of them, use this approach: Find the average correlation of A and B with the rest of the variables. Whichever variable has the higher average in comparison with other variables, remove it. Alternatively, you can use penalized regression methods such as lasso, ridge, elastic net, etc.

4. You can do variable selection based on p values. If a variable shows p value > 0.05, we can remove that variable from model since at $p > 0.05$, we'll always fail to reject null hypothesis.

Linear Regression Model

Important Point 1 : Box Cox Transformation of Dependent Variable can solve problem of non-linearity, non-normality of error and heteroscedasticity.

Important Point 2 : RMSE for Training vs Test Sample

The RMSE for your training and your test sets should be very similar if you have built a good model. If the RMSE for the test set is much higher than that of the training set, it is likely that you've badly over fit the data, i.e. you've created a model that tests well in sample, but has little predictive value when tested out of sample.

Linear Regression Model

Important Point 3 : Transformation Rules

The specific transformation used depends on the extent of the deviation from normality.

1. If the distribution differs moderately from normality, a square root transformation is often the best.
2. A log transformation is usually best if the data are more substantially non-normal.
3. An inverse transformation should be tried for severely non-normal data..