greatlearning

Predictive Modelling – Linear Regression

Linear Regression TOC

1. Introduction to Linear regression
    a. What is linear regression
    b. How is it built
    c. Measuring model accuracy
2. Structure of Linear Regression model
    a. coefficients
    b. Intercept
    c. error
3. OLS (Ordinary Least Square Method)
4. Gradient Descent method
5. Multivariate Linear Regression
6. Hypothesis Testing in Linear Regression
7. Assumptions of Linear Regression
    1. Testing for violations of the assumptions
8. Advantages and disadvantages of Linear Regression
9. Applications of Linear Regression

Introduction To Linear Regression

greatlearning

Introduction to Linear Regression Models (What is it?) -

1. The term "regression" generally refers to predicting a real number.

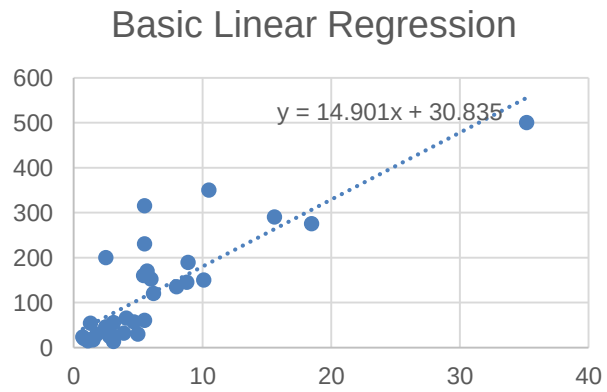
2. The term "linear" in the name “linear regression” refers to the fact that the method models data with linear combination of the explanatory variables.

3. A linear combination is an expression where one or more variables are scaled by a constant factor and added together.

4. In the case of simplest linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

   Dependent variable value =(weight ∗ independent variable) + constant

5. It is the straight line in the scatter plot of the variables



Introduction to Linear Regression Models (How is it built?) -

1. Before we create a linear model, we need to ensure the independent variable influences the dependent variable. The influence is technically called correlation

2. Mathematically correlation between two variables indicates how closely their relationship follows a straight line. By default we use Pearson's correlation which ranges between -1 and +1. This is represented by the character 'r'

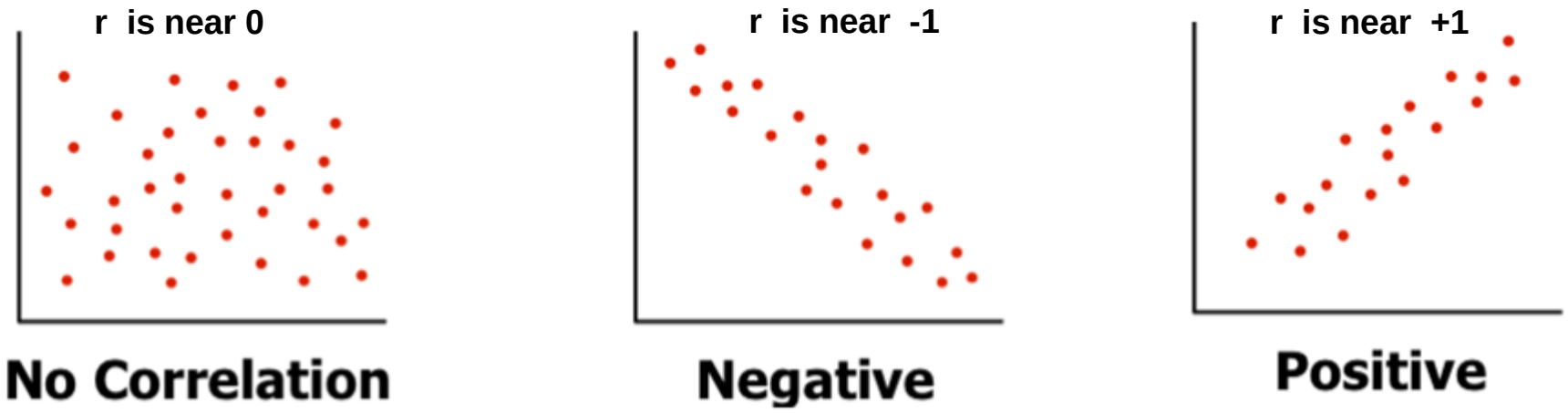
3. Correlation of extreme possible values of -1 and +1 indicate a perfectly linear relationship between X and Y whereas a correlation of 0 indicates absence of linear relationship
    a. When r value is small, one needs to test whether it is statistically significant or not to believe that there is correlation or not
    b. When r value is close to 0 it means there is no linear relationship between y and X. It does not mean there is no relation between them
    c. Negative correlation indicates inverse relation i.e. when one increases other decreases while positive correlation indicates direct relation i.e. both variables increase and decrease together

Introduction to Linear Regression Models (How is it built?) -

greatlearning

d. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

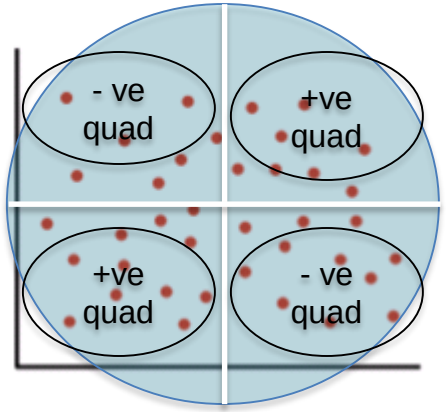


e. Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! **

** if the relation between Y and X is non-linear for e.g. y = X^2 , we can transform the independent variable for e.g. K = X^2 and build linear model using y and K i.e. y = K

Introduction to Linear Regression Models (How is it built?) -

f. When r is close to 0, indicates distribution of data points that look like a cloud. A value close to +1 or -1 indicate a strong linear trend in the distribution

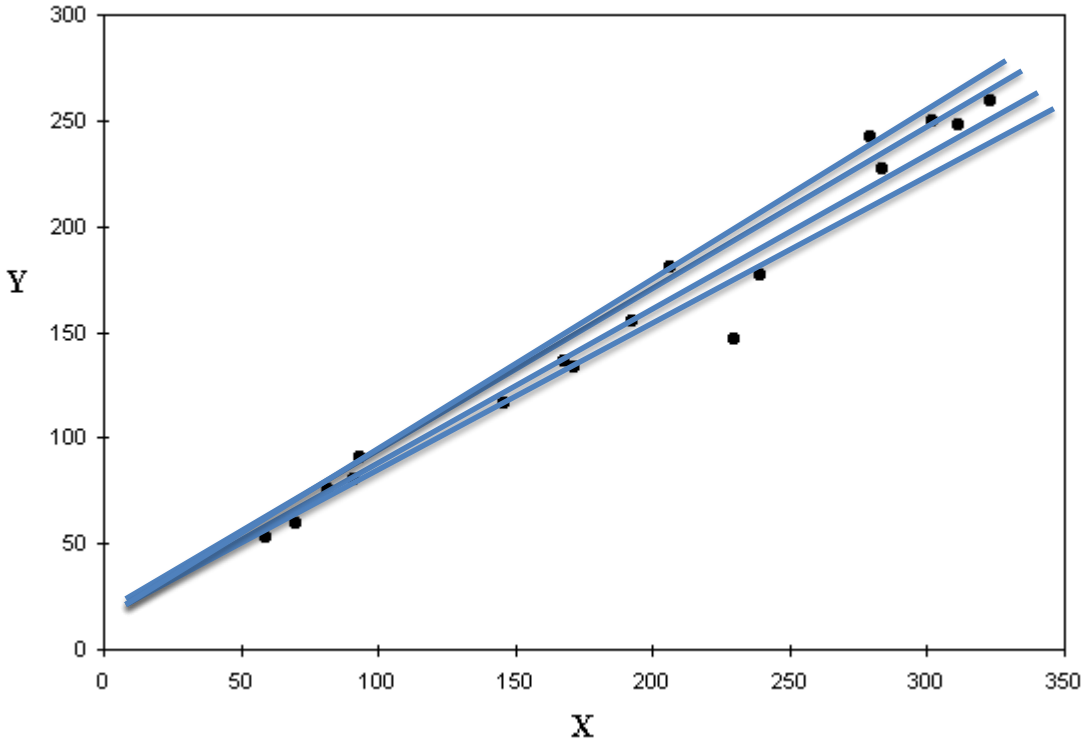


$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = 0$$

Introduction to Linear Regression Models (How is it built?) -

g. Given Y = f(x) and the scatter plot shows apparent correlation between X and Y Let's fit a line into the scatter which shall be our model

h. But there are infinite number of lines that can be fit in the scatter. Which one should we consider as the model?

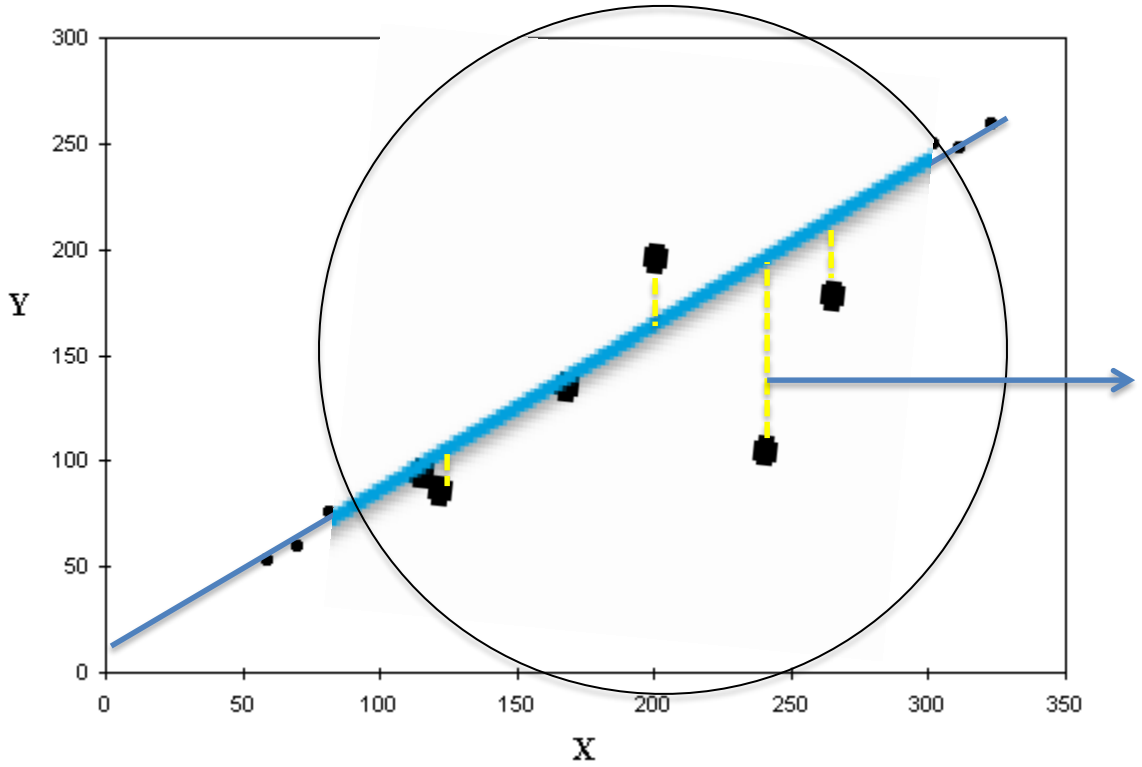


i. This and many other algorithms use gradient descent or variants of gradient descent method for finding the best model

j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

Accuracy of Linear Regression Models -

1. The line that represents the model may not touch all the points in the scatter plot
2. The vertical distance between a point and the line (shown in yellow) is the error in prediction of the model
3. The line which gives least sum of squared errors across all the data points put together is considered as the best line



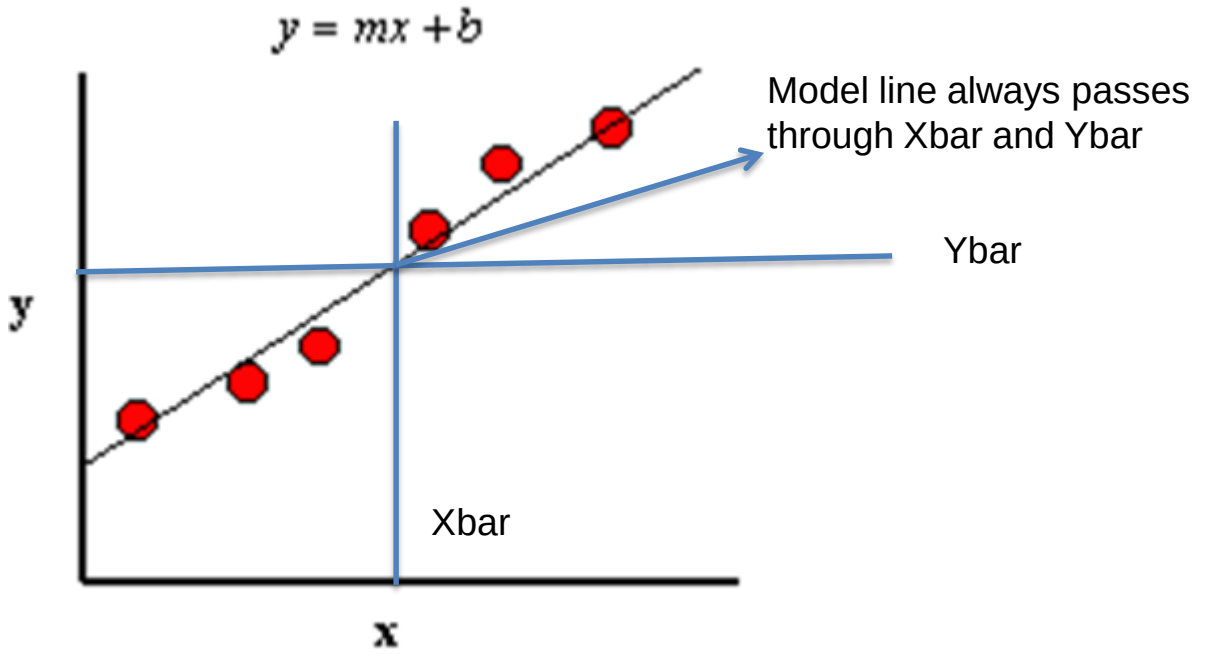
Error = (T – (mx + C)

Sum of all errors can cancel out and give 0

We square all the errors and sum it up. That line which gives us least sum of squared errors is the best fit

Accuracy of Linear Regression Models -

4. The best fit line will always go through that point in the features space where the Xbar (blue vertical line) and ybar (blue horizontal line) meet
5. Coefficient of determinant – determines the fitness of a linear model. The closer the points get to the line, the R^2 (coeff of determinant) tends to 1, the better the model is



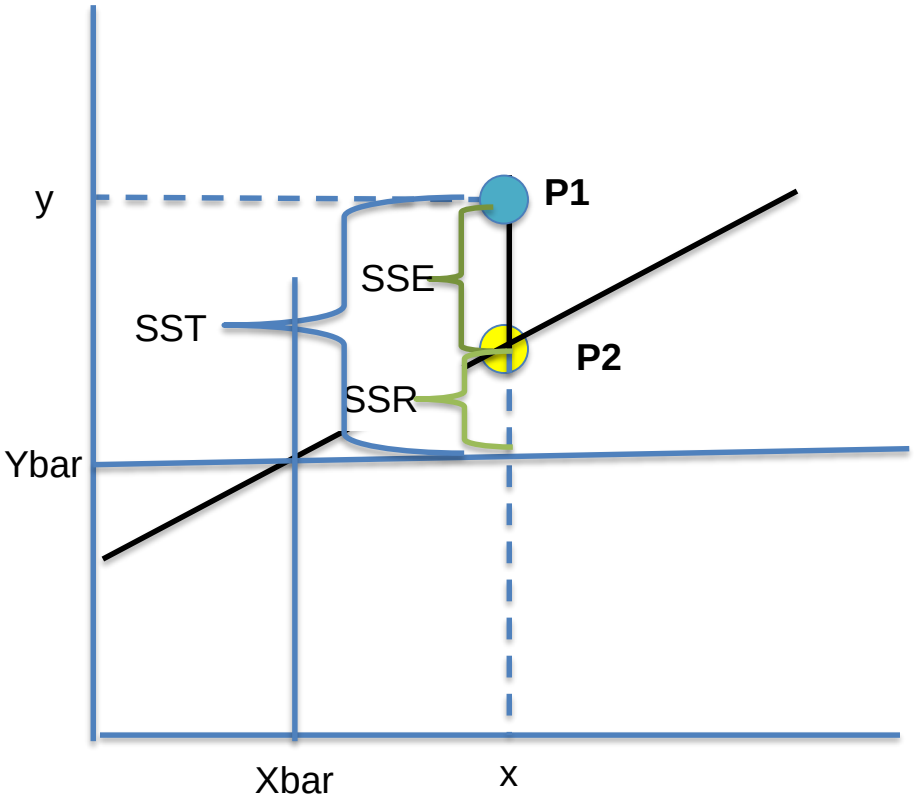
Accuracy of Linear Regression Models -

5. Coefficient of determinant (Contd…)

I. There are a variety of errors for all those points that don't fall exactly on the line.

II. It is important to understand these errors to judge the goodness of fit of the model i.e. How representative the model is likely to be in general

III. Let us look at point P1 which is one of the given data points and associated errors due to the model



a. P1 – Original y data point for given x

b. P2 - Estimated y value for given x

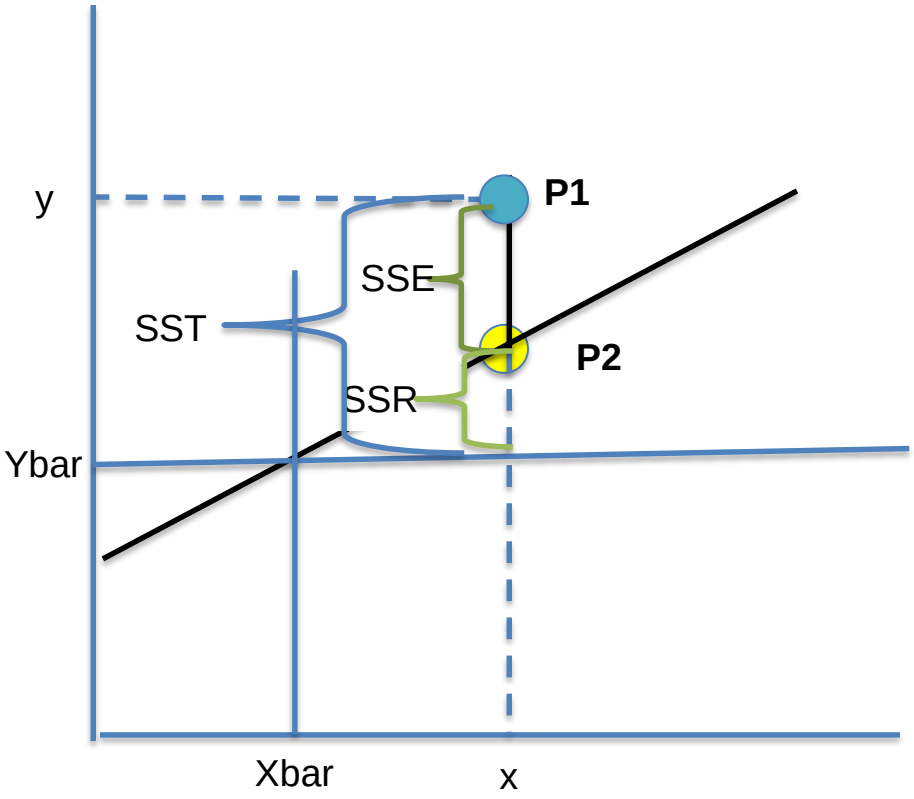
c. Ybar – Average of all Y values in data set

d. SST – Sum of Square error Total (SST) Variance of P1 from Ybar (Y – Ybar)^2

e. SSR - Regression error (p2 – ybar)^2 (portion SST captured by regression model)

f. SSE - Residual error (p1 – p2)^2

Accuracy of Linear Regression Models -



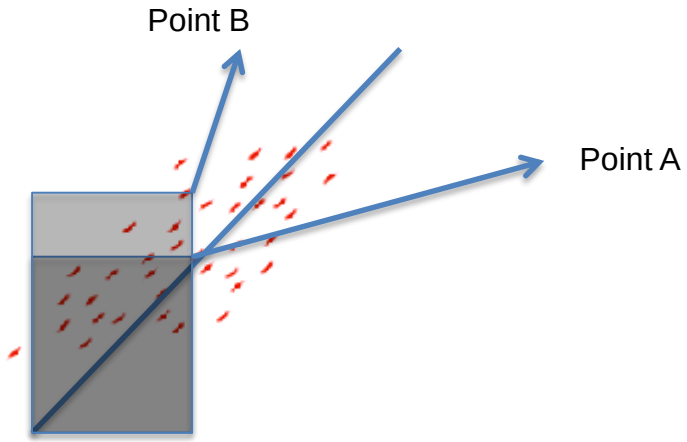
5. Coefficient of determinant (Contd…)

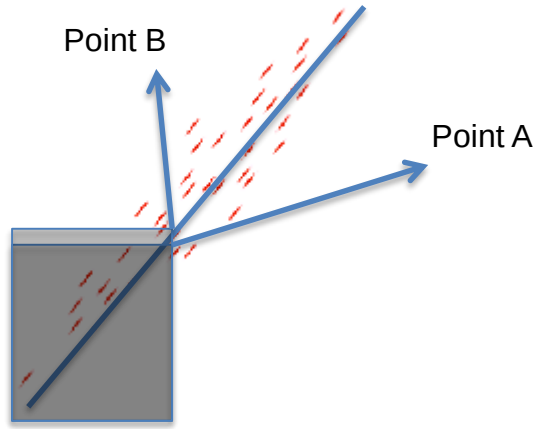
   1. That model is the most fit where every data point lies on the line. i.e. SSE = 0 for all data points
   2. Hence SSR should be equal to SST i.e. SSR/SST should be 1.
   3. Poor fit will mean large SSE. SSR/SST will be close to 0
   4. SSR / SST is called as r^2 (r square) or coefficient of determination
   5. r^2 is always between 0 and 1 and is a measure of utility of the regression model

Note: SS in all the terms stand for Sum Squared. In the diagram only one point is shown, to explain the concept. However, these terms make sense only when more than one data points are considered.

Accuracy of Linear Regression Models -

5. Coefficient of determinant (Contd…) -





In case of point “A”, the line explains the variance of the point

Whereas point “B” the is a small area (light grey) which the line does not represent.

%age of total variance that is represented by the line is coeff of determinant

Structure of Linear Regression

Structure of Linear Models

$Y = m_1X_1 + m_2X_2 + ... + m_nX_n + C + e$

Y = Dependent / target / predicted variable

X_i = Indepenedent / predictor variable

m_i = coefficients for the ith independent / predictor variable

C = constant / intercept / bias

e = residual error / unexplained variance / difference between actual and prediction

Ordinary Least Square Method

OLS (Ordinary Least Squares)

1. Also known as “Linear least squares”, estimates the parameters (coefficients, bias) in a linear regression model by minimizing the sum of the squared errors /residuals.
2. This method defines the linear regression model as that line which while passing through the distribution of data points, minimizes the sum of the squared differences between the observed values and the predicted value
3. The combination of coefficients and bias that minimize the sum of squared errors can be found using simple algebraic operations
4. Let Yp be the predicted value of Y (the actual value) for a given independent variable value of X
5. Let m and c be the coefficient and the bias respectively
6. The linear model thus is $Yp = mX + c$
7. Error in prediction = $Y - Yp = Y - (mX + c)$
8. We have to find the m and c (best m and c) which minimizes the error
   a. Finding the best m – (corr (X,Y) * stddev_Y) / stddev_X
   b. Finding the best c - Ybar – m_best * Xbar
9. This method is not scalable i.e. with increasing independent variables and increasing data points, this method will take long to find the best fit line. Hence, in data science we use another method called “Gradient Descent”

Gradient Descent method

1. In this method the error is represented in squared form i.e. $E = (Y_{expected} - Y_{pred})^2$
2. Expanding Ypred, $E = ((Y_{expected} - (mX + C))^2$
3. Thus, E is a function of m and c given $Y_{expected}$ and X come from data
4. The E function being quadratic (raised to power of 2) when plotted against m and c, will acquire a parabolic shape
5. This guarantees an absolute minima i.e. there will be a unique combination of m and c which will deliver the least error. Let his be the best m and best c
6. Starting from some random m and c, the Gradient Descent method will automatically discover the best m and best c using a mathematical technique called “Partial Derivatives”
7. This method can be applied with any number of independent variables. It will be faster than the algebraic method

Multivariate Linear Regression

Multivariate Linear Regression

1. When more than two predictor variables are used to predict the value in the dependent variable

2. The structure of the model remains same but gets extended to include all the variables instead of just one as in simple linear regression
   a. $Y = m_1X_1 + m_2X_2 + \ldots + m_nX_n + c + e$

3. Geometrically, the line in simple linear regression model is replaced with a plane (for two predictor variables) and by a hyper plane (planes in higher than three dimensions) to express the relationship between dependent and independent variables

4. The predictor variables are expected to be independent of one another i.e not correlate amongst themselves

Mpg_Linear+Regression_statsmodel.ipynb

Multivariate Linear Regression

1. Model coefficients for car-mpg.csv dataset

Attributes	Coefficient	Interpretation
cyl	1.863717834	one unit increase in cyl, mpg increases by 1.86 units
disp	0.010066051	one unit increase in disp, mpg increases by 0.01 units
hp	-0.039229006	one unit increase in hp decreases mpg by 0.03 units
wt	-0.006414997	one unit increase in wt decreases mpg by 0.006 units
acc	0.011723809	one unit increase in acc increases mpg by 0.01
yr	0.758818485	one unit increase in year of manufacture increases mpg by .75 units
car_type	6.626521339	Automatic car type increases mpg by 6.6 units
Intercept	-26.69336013	Meaning less.

2. Model structure - mpg = 1.86 cyl + 0.01 disp – 0.03 hp - 0.006wt + 0.01 acc + 0.75 yr + 6.6 cr_type – 26.6

3. The intercept of -26.6 is meaningless and can be eliminated if we scale the data using Zscore (centring data). This does not impact the accuracy but may change the coefficients.

Mpg_Linear+Regression_statsmodel.ipynb

Hypothesis Testing In Linear Regression

Hypothesis Testing For Linear Regression

1. The big question one needs to ask is, are these coefficients really reflecting the relation between the target variable and the independent variable or are they by statistical chance.
2. Consider the coefficient for the acceleration variable, it is positive 0.01! What it means is, for every increase in acceleration, mpg increases by 0.01 units. Which is absurd!
3. To establish the reliability of the coefficients, we need hypothesis testing
4. The Null hypothesis (H0) claims there is no relation between mpg and any of the variables. That means the coefficient is 0 in the universe
5. Assuming H0 to be true, what is the probability of finding the coefficients found in the sample if the sample is drawn from that universe in which H0 is true
6. At 95% confidence level if the p value is < .05, we reject the H0 i.e. probability of finding these coefficients in sample if they are 0 in the universe is very low
7. If p value is >=.05, we do not have sufficient evidence in the data to reject the H0 and hence we do not reject H0. We believe H0 is likely to be true in the universe

Hypothesis Testing For Linear Regression

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.826
Model:	OLS	Adj. R-squared:	0.821
Method:	Least Squares	F-statistic:	182.9
Date:	Fri, 20 Dec 2019	Prob (F-statistic):	1.41e-98
Time:	13:51:35	Log-Likelihood:	-725.17
No. Observations:	278	AIC:	1466.
Df Residuals:	270	BIC:	1495.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-26.6934	5.597	-4.769	0.000	-37.713	-15.674
cyl	1.8637	0.517	3.606	0.000	0.846	2.881
disp	0.0101	0.009	1.123	0.262	-0.008	0.028
hp	-0.0392	0.016	-2.420	0.016	-0.071	-0.007
wt	-0.0064	0.001	-7.865	0.000	-0.008	-0.005
acc	0.0117	0.114	0.103	0.918	-0.212	0.236
yr	0.7588	0.060	12.668	0.000	0.641	0.877
car_type	6.6265	1.041	6.364	0.000	4.577	8.677

Omnibus:	35.838	Durbin-Watson:	2.082
Prob(Omnibus):	0.000	Jarque-Bera (JB):	68.579
Skew:	0.693	Prob(JB):	1.28e-15
Kurtosis:	5.000	Cond. No.	8.64e+04

1. R-Squared metric is not reliable as it is does not take into account spurious correlations
2. Adjusted R-squared metric accounts for the spurious correlations. It decreases when we include attributes into the model that are weak or poor predictors of Y
3. The statistical analysis is based on T distribution which as mean of 0
4. Coeff = Std_error * t where std_error is standard deviation and t is zscore in T distribution
5. P is the conditional probability given H0 is true
6. Atrributes such as disp, acc have P value > 0.05 and hence statistically their coefficients are not reliable
7. Overall, model P value is lower than 0.05 which means model is reliable after eliminating the useless attributes

Assumptions In Linear Regression

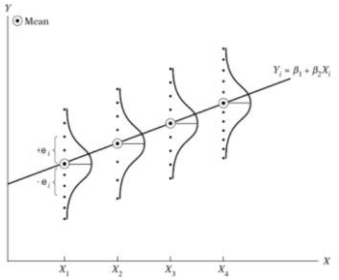
Regression Model Assumptions

Assumption 1

1. Population Linear Regression Model is linear in the parameters though it may not be linear in the variables i.e. the slope coefficients are always raised to power 1. The variables may be raised to any power. The regression model thus, takes the form $y_i = m_iX_i + C + e_i$ (C is intercept, mi is coefficient, X_i is variable, e_i is disturbance

Assumption 2

1. The mean value of error e_i is zero. Given the value of X_i, the means or expected value of the random disturbance $E(e_i \mid X_i) = 0$ i.e. $E(e_i) = 0$
2. Population of y corresponding to a given X_i is distributed around its mean value, implies no specification bias / error in the model indicating that the model is correctly specified.



Assumption 3

1. The error in prediction of each trial is independent of the value of X.
2. Error term e_i, represents the impact of the variables not considered for the model

Regression Model Assumptions

Assumption 4

1. Homoscedasticity or Constant Variance of e_i, the variance of the error / disturbance is the same regardless of the value of X
2. Var(ui) = E[e_i - E(e_i | X_i)]^2 = E(e_i^2 | X_i) because of assumption 3 (E(e_i) = 0).
3. = E(e_i^2) for a given X_i = constant variance (representation for homoscedasticity)

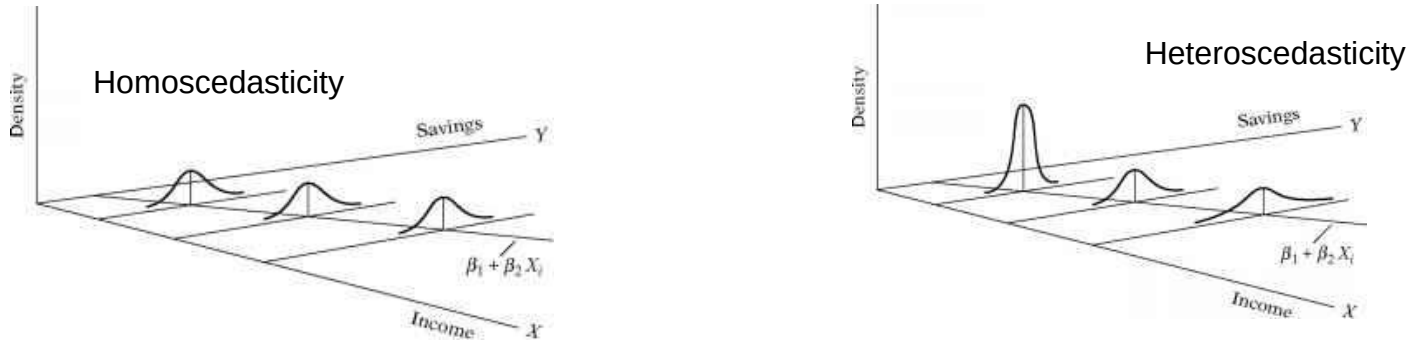


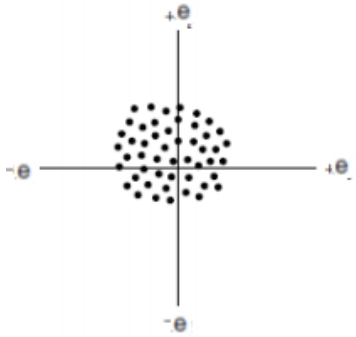
Image source: https://www.rhayden.us/regression-models/the-nature-of-heteroscedasticity.html

4. The likelihood that the y observations coming from the population with X = X_i would be closer to the population regression function than those coming from the populations corresponding to X = X_2, X = X_3 and so on. The reliability of predicted Y will fall
5. By invoking Assumption 4, we stress equal importance to all y values corresponding to different values of X

Regression Model Assumptions

Assumption 5

1. No autocorrelation between disturbances e_i. Given any two X values, X_i and X_j (i <> j), the correlation between any two e_i and e_j is zero i.e. no serial or auto correlation
2. This assumption is justified when time is not an attribute i.e.



Assumption 6

1. The number of observations n must be greater than the number of parameters to be estimated. In data science parlance, the depth should be much greater than breadth i.e. number of records much larger than the number of columns to avoid curse of dimensionality situation.

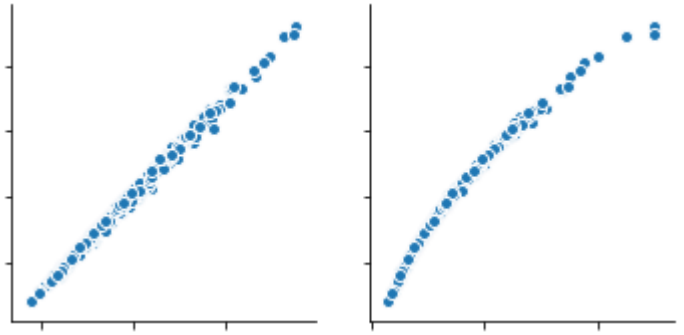
Assumption 7

1. The X should have variance. The values should not be constant. In Data Science parlance, X should have variance. Further, the outliers should not exist

Regression Model Assumptions

Assumption 8

1. There no perfect collinearity between the predictor variables X
2. In case of perfect collinearity, the scatter plot will be line
3. Most often we come across less than perfect collinearity



Assumption 9

1. The model is correctly specified i.e. neither overfit or underfit

Assumption 10

The stochastic term e_i is normally distributed. The error term 'e' follows the normal distribution with zero mean and (constant) variance

$$e_i \sim N(0, \sigma^2)$$

where the symbol ~ means distributed as and N stands for the normal distribution, the terms in the parentheses representing the two parameters of the normal distribution, namely, the mean and the variance. If this assumption is violated, the statistical tests such as t, and F in regression may not be valid.

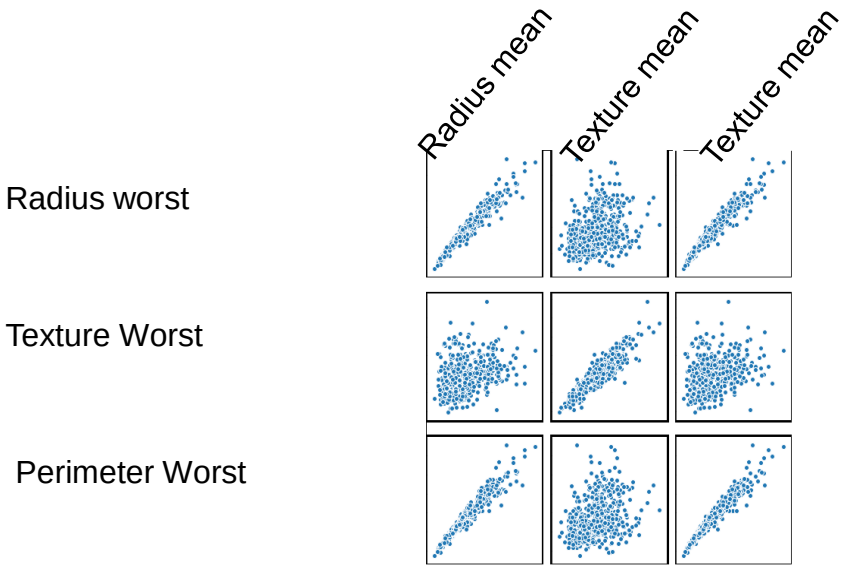
Note: All the assumptions pertain to population regression function only.

greatlearning

Multi-Collinearity

Multi Collinearity – What is it?

Multicollinearity is that situation where the independent variables in the linear model are not truly independent i.e. they are correlated. For e.g. in the Wisconsin Breast Cancer dataset the first three attributes “radius mean”, “perimeter mean” and “texture mean” is shown below. The first two are strongly correlated



Multi Collinearity – Types of multicollinearity

1. **Structural multicollinearity**: This type occurs when we create features from existing features and build a model using all of the features. For example, using “Radius” and “Area” as two variables. When features are generated, ensure the generated feature and the original features do not strongly correlate, if they do, you may want to drop the original feature as long as the generated feature contains all the information from the original

2. **Data multicollinearity**: This type of multicollinearity is an artifact of the data itself. The nature of the variables is such that they correlate. For e.g. in auto-mpg.csv, the columns “weight” and “horsepower” of a car will correlate positively. In case there are such correlating variables in the data, they may be combined into a composite variable using techniques such as PCA

3. The problem with having multicollinearity is in the inability to understand how one variable influences the target. There is no way to estimate separate influence of each variable on target. Thus no way to estimate the partial regression coefficients

Multi Collinearity (Contd…)

4. If multicollinearity is perfect, the regression coefficients of X variables are indeterminate and their standard errors are infinite

5. If multicollinearity is less than perfect, the regression coefficients, although determinate, posses large standard errors, which means the coefficients cannot be estimated with confidence

6. High degree of multicollinearity will not take away the property of being best unbiased linear estimators. It violates none of the regression assumptions. The only problem is that it will result in hard to determine coefficients with small standard errors

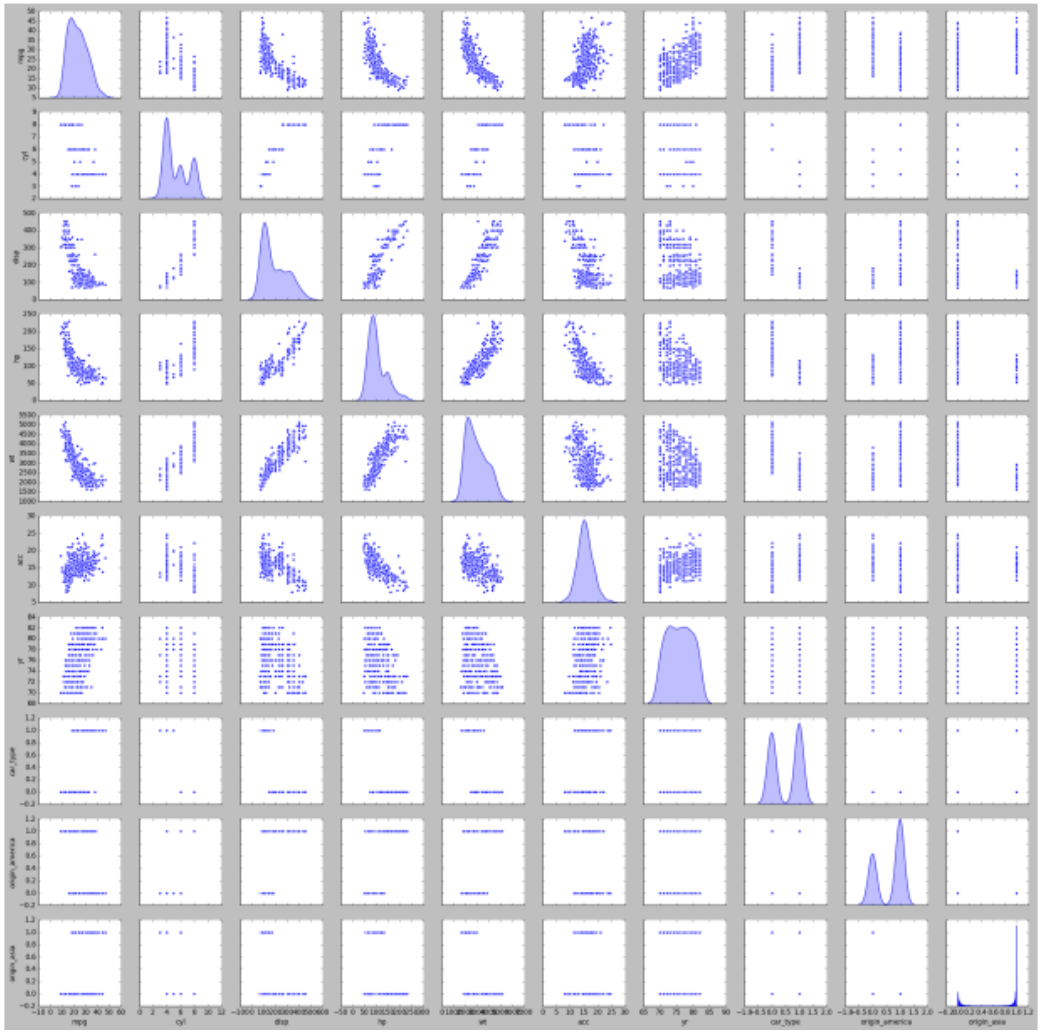
Multi Collinearity – Testing for multicollinearity with Variation Inflation Factor (VIF)

1. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation
    a. VIF = $1 / 1 - r^2$
    b. As the collinearity between variables increase, r^2 increases, the denominator approaches 0 and as a result VIF approaches infinity

2. VIFs start at 1 and have no upper limit
    a. A value of 1 indicates that there is no correlation between this independent variable and any others
    b. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures
    c. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

Testing for multicollinearity with Variation Inflation Factor (VIF)

Car-mpg.csv dataset has many attributes that correlate with one another.



```
from statsmodels.stats.outliers_influence import
variance_inflation_factor
vif = [variance_inflation_factor(X.values, ix) for ix in
range(X.shape[1])]
i=0
for column in X.columns:
    if i < 11:
        print (column ,"--->",  vif[i])
        i = i+1
```

```
cyl ---> 172.09167529137474
disp ---> 87.05808335183303
hp ---> 71.23983108333236
wt ---> 139.1665144189037
acc ---> 69.82068667385671
yr ---> 166.95012233353933
car_type ---> 12.993508077923245
```

Mpg_Linear+Regression_statsmodel.ipynb

Testing For Linear Regression Assumptions

Testing for violations of the assumptions

1. Residual Vs fitted plots

When conducting a residual analysis, a "residuals versus fits plot" is the most frequently created plot. It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

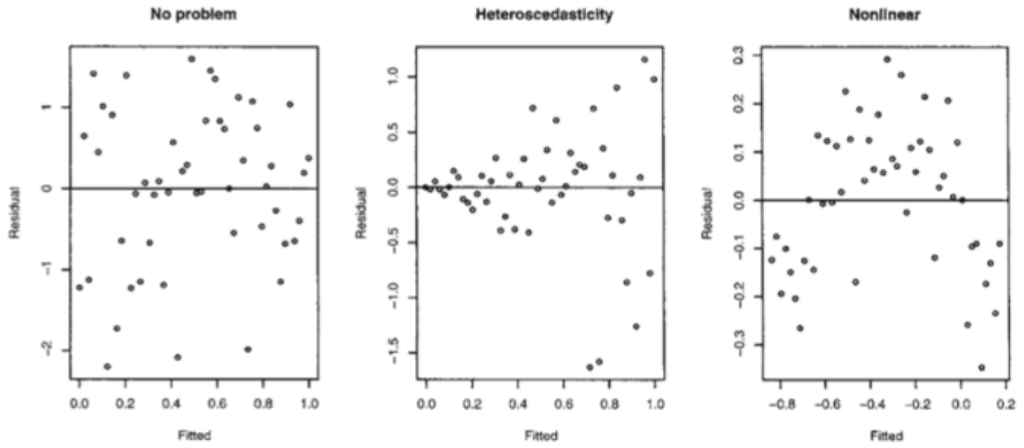


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

Applications of Linear Regression

Applications of Linear Regression-

Retail –

1. Customer lifetime value analysis – Factoring in customer purchase characteristics to estimate customer value in near future. Model based on history of purchase transactions, customer attributes
2. Predicting optimal price – factor in characteristics of the object to predict optimal sale price. For e.g. real estate pricing, car re-sale price

Banking & Finance –

1. Portfolio profit / loss forecasting – use economic factors such as GDP growth rate, Stock market performance, national budgets, national level policies etc
2. Estimating growth in customer accounts given the economic parameters, sales strategies and promotions etc

Education –

1. Predict student scores given the attributes such as number of classes attended, past performance, family support etc.
2. Predict number of seats required in various courses next academic year

Applications of Linear Regression- (Contd…)

Medical –

1. Predict number of hospital beds required in near future given current demand, season and other social factors
2. Predict health care cost for insured customers given their attributes such as age, race, gender, previous claims

Manufacturing –

1. Demand estimation / capacity estimation – predict demand and capacity for a defined period given the socio economic factors, season

Linear Regression Model -

Advantages –

1. Simple to implement and easier to interpret the outputs coefficients
2. Works well even when the relation between independent variable and dependent variable are not linear if we transform the variables

Disadvantages -

1. Assumes a linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them
2. Outliers can have huge effects on the regression
3. Linear regression assume independence between attributes
4. Linear regression looks at a relationship between the mean of the dependent variable and the independent variables.
5. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables

greatlearning

Thank You

Fine Tuning Linear Regression Models

Regularising Linear Models (Shrinkage methods)

1. Sometimes we have to engineer features from existing features to improve accuracy. For e.g. we may create polynomial features using Sklearn polynomial feature generators

2. Though the polynomial features will improve model accuracy in training, the improvement may come at the risk of curse of dimensionality leading to an over fit model. To prevent curse of dimensionality situation, we employ the regularization techniques of Ridge or Lasso

3. Shrinkage methods attempt to shrink the coefficients of the attributes and lead us towards simpler yet effective models. The two shrinkage methods are :

    1. Ridge regression is similar to the linear regression where the objective is to find the best fit surface. The difference is in the way the best coefficients are found. Unlike linear regression where the optimization function is SSE, here it is slightly

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

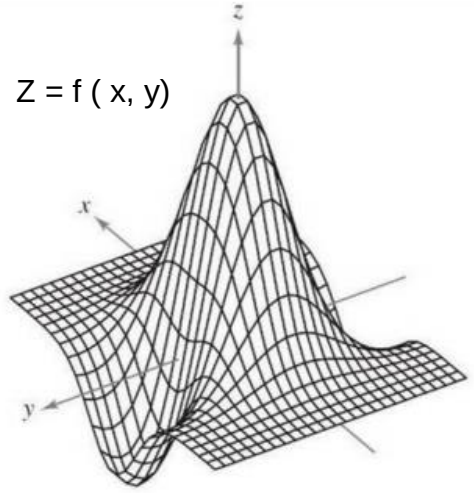
Linear Regression cost function

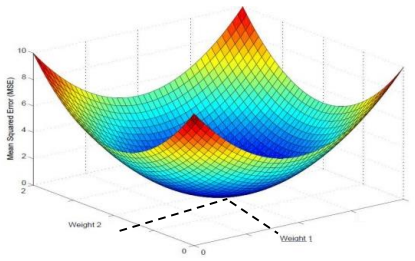
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Ridge Regression with additional term in the cost function

    2. The term is like a penalty term used to penalize large magnitude coefficients when it is set to a high number, coefficients are suppressed significantly. When it is set to 0, the cost function becomes same as linear regression cost function

Regularising Linear Models (Shrinkage methods)

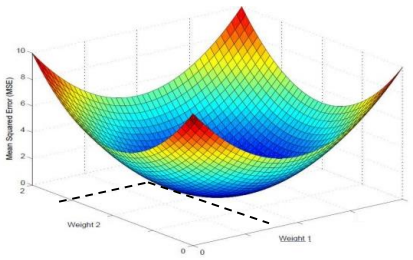




$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2=0$$

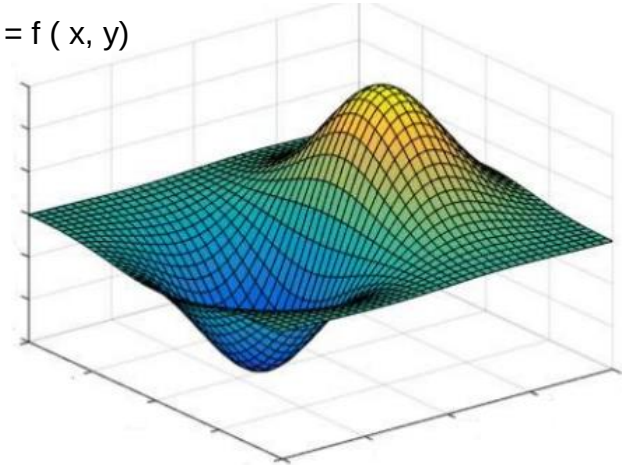
1. Curse of dimensionality results in large magnitude coefficients which results in a complex undulated surface / model.
2. This complex surface has the data points occupying the peaks and the valleys
3. The model gives near 100% accuracy in training but poor result in testing and the testing scores also vary a lot from one sample to another.
4. The model is supposed to have absorbed the noise in the data distribution!
5. Large magnitudes of the coefficient give the least SSE and at times SSE = 0! A model that fits the training set 100%!
6. Such models do not generalize

Regularising Linear Models (Shrinkage methods)



$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Z = f (x, y)



1. In Ridge Regression, the algorithm while trying to find the best combination of coefficients which minimize the SSE on the training data, is constrained by the penalty term

2. The penalty term is akin to cost of magnitude of the coefficients. Higher the magnitude, more the cost. Thus to minimize the cost, the coefficient are suppressed

3. Thus the resulting surface tends to be relatively much more smoother than the unconstrained surface. This means we have settled for a model which will make errors in the training data

4. This is fine as long as the errors can be attributed to the random fluctuations i.e. because the model does not absorb the random fluctuations in the data

5. Such model will perform equally well on unseen data i.e. test data. The model will generalize better than the complex model

Regularising Linear Models (Shrinkage methods)

Impact of Ridge Regression on the coefficients of the 56 attributes

```
Ridge model: [[ 0. 3.73512981 -2.93500874 -2.13974194 -3.56547812 -1.28898893 3.01290805
2.04739082 0.0786974 0.21972225 -0.3302341 -1.46231096 -1.17221896 0.00856067 2.48054694
-1.67596093 0.99537516 -2.29024279 4.7699338 -2.08598898 0.34009408 0.35024058 -0.41761834
3.06970569 -2.21649433 1.86339518 -2.62934278 0.38596397 0.12088534 -0.53440382 -1.88265835
-0.7675926 -0.90146842 0.52416091 0.59678246 -0.26349448 0.5827378 -3.02842915 -0.36548074
0.5956112 -0.15941014 0.49168856 1.45652375 -0.43819158 -0.20964198 0.77665496 0.36489921
 -0.4750838 0.3551047 0.23188557 -1.42941282 2.06831543 -0.34986402 -0.32320394 0.39054656 0.06283411]]
```

Large coefficients have been suppressed, almost close to 0 in many cases.

Ref: Ridge_Lasso_Regression.ipynb

Regularising Linear Models (Shrinkage methods)

1. Lasso Regression is similar to the Ridge regression with a difference in the penalty term. Unlike Ridge, the penalty term here is raised to power 1. Also known as L1 norm.

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

2. The term λ continues to be the input parameter which will decide how high penalties would be for the coefficients. Larger the value more diminished the coefficients will be.

3. Unlike Ridge regression, where the coefficients are driven towards zero but may not become zero, Lasso Regression penalty process will make many of the coefficients 0. In other words, literally drop the dimensions

Regularising Linear Models (Shrinkage methods)

Impact of Lasso Regression on the coefficients of the 56 attributes

Lasso model: [0. 0.52263805 -0.5402102 -1.99423315 -4.55360385 -0.85285179 2.99044036 0.00711821 -0. 0.76073274 -0. -0. -0.19736449
0. 2.04221833 -1.00014513 0. -0. 4.28412669 -0. 0. 0.31442062 -0. 2.13894094 -1.06760107 0. -0. 0. 0. -0.44991392 -1.55885506 -0. -0.68837902 0.
0.17455864 -0.34653644 0.3313704 -2.84931966 0. -0.34340563 0.00815105 0.47019445 1.25759712 -0.69634581 0. 0.55528147 0.2948979 -0.67289549
0.06490671 0. -1.19639935 1.06711702 0. -0.88034391 0. -0.]

Large coefficients have been suppressed, to 0 in many cases, making those dimensions useless i.e. dropped from the model.

Ref: Ridge_Lasso_Regression.ipynb