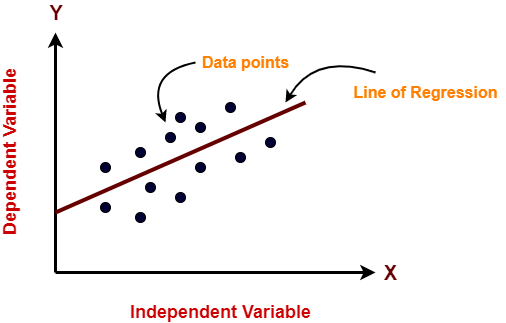
***Linear Regression***

**Aim**: Health insurance companies invest more on machine learning algorithms to forecast expenses on medical belonging to population. Example, chances of lung cancer is high among smokers than non-smokers. Though medical conditions are hard to predict and are rare among most of the conditions, some can be predicted like the example mentioned above. Insurance companies can impose higher premium on the people with forecasted ailments. This can bring their money back they spend on their recipients. Here, we are going to make a model which can predict medical care expenses using patient data by which premiums can be set.

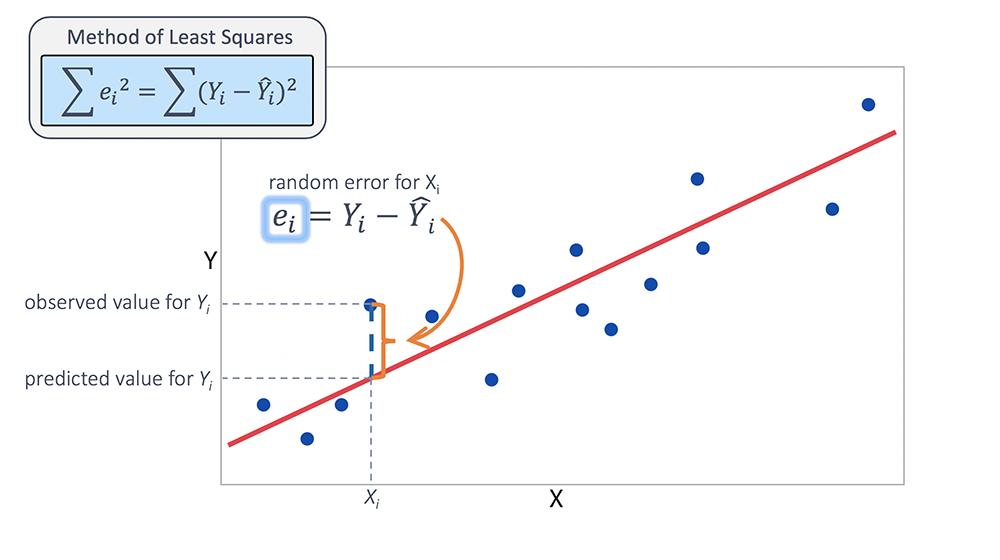
**Algorithm selected**: Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables. If there is single independent variable, it is called simple regression. If there are multiple independent variables, then it is multiple regression.

Linear regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression. On X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

Y=mX+c

,

Best fit line is for getting optimum m and c is obtained by OLS estimation. These are minimized such a way that sum of the squared errors, that is, the vertical distance between the predicted y value and the actual y value is low. Predicted y values are on the best fit line. These errors are known as residuals and error term is epsilon.



R generates best fit line for multiple points to make sure that there is minimum error rate.

1. **DATA COLLECTION**: For this, dataset containing medical expenses of patients enrolled for insurance plan in US were used using demographic statistics from USCB.

* insurance=read.csv("credit.csv")

*age sex bmi children smoker region expenses*

*1 19 female 27.9 0 yes southwest 16884.92*

*2 18 male 33.8 1 no southeast 1725.55*

*3 28 male 33.0 3 no southeast 4449.46*

*4 33 male 22.7 0 no northwest 21984.47*

*5 32 male 28.9 0 no northwest 3866.86*

This data set has 1338 instances and 7 features on characteristics of patient data like age(64 above are excluded),sex(male ,female), bmi(18.5-24.9),children(dependents for patients),smoker(either yes or no), region(divided into 4 categories),expenses(indicates their insurance amount)

**2.DATA EXPLORING**: On Regression model, we must specify the relationships between the features. They aren’t automatically detected in regression.

* str(insurance)

*'data.frame': 1338 obs. of 7 variables:*

*$ age : int 19 18 28 33 32 31 46 37 37 60 ...*

*$ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...*

*$ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...*

*$ children: int 0 1 3 0 0 0 1 3 2 0 ...*

*$ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...*

*$ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...*

*$ expenses: num 16885 1726 4449 21984 3867 ...*

Here expenses is dependent variable and remaining are the independent variables. Lets see the distribution of expenses using histogram and summary

* summary(insurance$expenses)

*Min. 1st Qu. Median Mean 3rd Qu. Max.*

*1122 4740 9382 13270 16640 63770*

* hist(insurance$expenses)

This shows that the data is right skewed (Mean>Median)

Chart, histogram

Description automatically generated

From this, we can figure out that majority of population has insurance plan ranging from 0 to 15000$. There are categorical variables in the data, these should be converted into dummies for regression analysis

**3.DATA PREPARATION:** Before preparing the model, relationship between independent and dependent variables should be determined. Correlation matrix provides pairwise correlations of the data.

* cor(insurance[c("age", "bmi", "children", "expenses")])

*age bmi children expenses*

*age 1.0000000 0.10934101 0.04246900 0.29900819*

*bmi 0.1093410 1.00000000 0.01264471 0.19857626*

*children 0.0424690 0.01264471 1.00000000 0.06799823*

*expenses 0.2990082 0.19857626 0.06799823 1.00000000*

No correlation in matrix is strong. There is weak positive correlation between age and BMI, moderate positive correlation between age and expenses, BMI and expenses, children, and expenses. These suggest that if age, BMI, children increase, expected cost increases

We can also visualize relationships using scatter plot matrix.The relationship between age and expenses displays several relatively straight lines, while the bmi versus expenses plot has two distinct groups of points.

Diagram, engineering drawing

Description automatically generated

Adding more information to the above plot using pairs.panels() in psych pakage

* install.packages("psych")
* library(psych)
* pairs.panels(insurance[c("age", "bmi", "children", "expenses")])

Chart

Description automatically generated

This plot has more added information like correlation coefficients, histogram pdf displayed, and correlation ellipse showing strength of correlation. More it is stretched, the stronger the correlation. More rounded it is, weaker the correlation

**4.TRAINING DATA:** For regression, we use lm () function from stats package in R. Using lm function, data is fit into regression model by specifying dependent and independent features as follows

* library(stats)
* ins\_model=lm(expenses ~ age + children + bmi + sex +smoker + region, data = insurance)
* ins\_model

*Coefficients:*

*(Intercept) age sexmale bmi children*

*-11941.6 256.8 -131.4 339.3 475.7*

*smokeryes regionnorthwest regionsoutheast regionsouthwest*

*23847.5 -352.8 -1035.6 -959.3*

Lm() automatically applied dummy function to create dummy variables for factor level features by ,making female non-smokers in the northeast region the reference group.

Intercept coefficient is the value obtained for expenses when rest all features are 0. Other coefficients indicate the increase in expenses for each increase in corresponding variable when other variables are held constant.

**5.EVALUATING THE MODEL:** Evaluating the model performance by checking how well the model fits our data. This is checked using summary() for our model

* summary(ins\_model)

*Residuals:*

*Min 1Q Median 3Q Max*

*-11302.7 -2850.9 -979.6 1383.9 29981.7*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) -11941.6 987.8 -12.089 < 2e-16 \*\*\**

*age 256.8 11.9 21.586 < 2e-16 \*\*\**

*sexmale -131.3 332.9 -0.395 0.693255*

*bmi 339.3 28.6 11.864 < 2e-16 \*\*\**

*children 475.7 137.8 3.452 0.000574 \*\*\**

*smokeryes 23847.5 413.1 57.723 < 2e-16 \*\*\**

*regionnorthwest -352.8 476.3 -0.741 0.458976*

*regionsoutheast -1035.6 478.7 -2.163 0.030685 \**

*regionsouthwest -959.3 477.9 -2.007 0.044921 \**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 6062 on 1329 degrees of freedom*

*Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494*

*F-statistic: 500.9 on 8 and 1329 DF, p-value: < 2.2e-16*

Here there are 3 sections

Residual summary: -Residual is actual value minus predicted value, the maximum error of 29981.7 indicates the model under-predicted expenses by nearly $30,000 for at least one observation. On the other hand, 50 percent of errors fall within the 1Q and 3Q values (the first and third quartile), so most predictions were between $2,850.90 and $1,383.90 of the actual value.

Coefficient summary: - p value is an estimate of the probability that the true coefficient is zero given the value of the estimate. Small p-values suggest that the true coefficient is very unlikely to be zero, which means that the feature is extremely unlikely to have no relationship with the dependent variable. p-values less than the significance level are considered statistically significant. Here, our model has several highly significant variables, and they seem to be related to the outcome in logical ways.

Standard Errors: -The multiple R-squared value (also called the coefficient of determination) provides a measure of how well our model explains the values of the dependent variable. If it is closer to 1 the model better explains the data. Since the R-squared value is 0.7494, we know that the model explains nearly 75 percent of the variation in the dependent variable. It is useful for comparing the performance of models with different numbers of explanatory variables using adjusted R square as it penalizes models with many independent variables.

The overall accuracy of the model is 74.94 (75%). Let’s try improving the model.

**6.IMPROVING THE MODEL:** Subject matter knowledge is important in performing regression as the feature selection is not automatic and user has to define the variables. There may not be linear relationships in all the cases some may deviate

Here there are few relations to be improved upon.

1. age and expenses may not be linearly related as age increases medical expenses may decrease due to government allowances. So, modeling them as polynomial relation by squaring the age feature value
2. BMI may not have impact on medical expenses when it is in normal range. It affects expenses only the BMI is above 30 on scale. So, taking this into consideration for our improvement by taking BMI values above 30.
3. So far, we have considered only individual effect of features on expenses. Adding interactions also helps in improving the performance. BMI above 30 and smoking yes creates relatively more impact on expenses than individual affect.

Considering all these points and regenerating a model.

* insurance$age2=insurance$age^2
* insurance$bmi30= ifelse(insurance$bmi >= 30, 1, 0)
* ins\_model2=lm(expenses ~ age + age2 + children + bmi + sex +bmi30\*smoker + region, data = insurance)
* summary(ins\_model2)

*Residuals:*

*Min 1Q Median 3Q Max*

*-17297.1 -1656.0 -1262.7 -727.8 24161.6*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 139.0053 1363.1359 0.102 0.918792*

*age -32.6181 59.8250 -0.545 0.585690*

*age2 3.7307 0.7463 4.999 6.54e-07 \*\*\**

*children 678.6017 105.8855 6.409 2.03e-10 \*\*\**

*bmi 119.7715 34.2796 3.494 0.000492 \*\*\**

*sexmale -496.7690 244.3713 -2.033 0.042267 \**

*bmi30 -997.9355 422.9607 -2.359 0.018449 \**

*smokeryes 13404.5952 439.9591 30.468 < 2e-16 \*\*\**

*regionnorthwest -279.1661 349.2826 -0.799 0.424285*

*regionsoutheast -828.0345 351.6484 -2.355 0.018682 \**

*regionsouthwest -1222.1619 350.5314 -3.487 0.000505 \*\*\**

*bmi30:smokeryes 19810.1534 604.6769 32.762 < 2e-16 \*\*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 4445 on 1326 degrees of freedom*

*Multiple R-squared: 0.8664, Adjusted R-squared: 0.8653*

*F-statistic: 781.7 on 11 and 1326 DF, p-value: < 2.2e-16*

Adjusted R squared value increased from 75% to 86.5% and also the age^2 and BMI30 are statistically significant which means our model is good

**Conclusion:** Using the lm() we were able to arrive at 75% accuracy at first. Improving the model by considering some of the specific interaction helped us increase the accuracy to 86.5%. Therefore, subject matter knowledge is very import to determines the relationship between independent features and on dependent features.

**Limitations of Linear Regression:**

* As subject knowledge is very important, it takes more time for beginner to understand the data and their relationships
* linear regression takes linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them. Sometimes this is wrong as is our case.
* These compares mean of dependent variables with independent variables. Mean is not a complete description of a single variable. We can consider medians and quantile ranges
* Linear regression assumes that the data are independent. Excluding the data having interactions(dependencies) might have impact on the accuracies of model.