MULTIPLE REGRESSION MODELLING

AIM: To built a multiple regression model of a given data.

R CODE AND OUTPUT:

```
> mydata2=read.table(file.choose(),header=T,sep=",")
> head(mydata2)
  patient_id risk Age Pressure Smoker Diabetes Fam_his
1        201   28  59      196     No       No      No
2        202   28  58       98     No       No      No
3        203   59  66      166     No       No      No
4        204   65  67      163     No       No     Yes
5        205   64  78      120     No       No     Yes
6        206   59  57      152     No      Yes      No
> mydata2$Smoker_new <-ifelse(mydata2$Smoker==c("Yes"),1,0)
> mydata2$Diabetes_new <-ifelse(mydata2$Diabetes==c("Yes"),1,0)
> mydata2$Fam_his_new <-ifelse(mydata2$Fam_his==c("Yes"),1,0)
> head(mydata2)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new
1        201   28  59      196     No       No      No          0            0           0
2        202   28  58       98     No       No      No          0            0           0
3        203   59  66      166     No       No      No          0            0           0
4        204   65  67      163     No       No     Yes          0            0           1
5        205   64  78      120     No       No     Yes          0            0           1
6        206   59  57      152     No      Yes      No          0            1           0

> mymodel2=lm(risk ~ Age + Pressure + Smoker_new + Diabetes_new + Fam_his_new,data=mydata2)
> summary(mymodel2)

Call:
lm(formula = risk ~ Age + Pressure + Smoker_new + Diabetes_new +
    Fam_his_new, data = mydata2)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7431  -7.4556   0.9263   5.4507  16.7411

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.28452   18.36081   1.649    0.108
Age           0.14159    0.22693   0.624    0.537
Pressure      0.01584    0.05058   0.313    0.756
Smoker_new   21.44273    4.59084   4.671 4.57e-05 ***
Diabetes_new 13.00869    2.93768   4.428 9.34e-05 ***
Fam_his_new  18.50377    3.56991   5.183 9.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.232 on 34 degrees of freedom
Multiple R-squared:  0.7643,    Adjusted R-squared:  0.7296
F-statistic: 22.05 on 5 and 34 DF,  p-value: 8.605e-10

> attach(mydata2)
> mydata2$predicted_value=41.227+23.661*Smoker_new+13.061*Diabetes_new+19.607*Fam_his_new
> head(mydata2)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new predicted_value
1        201   28  59      196     No       No      No          0            0           0          41.227
2        202   28  58       98     No       No      No          0            0           0          41.227
3        203   59  66      166     No       No      No          0            0           0          41.227
4        204   65  67      163     No       No     Yes          0            0           1          60.834
5        205   64  78      120     No       No     Yes          0            0           1          60.834
6        206   59  57      152     No      Yes      No          0            1           0          54.288
```

```
> mydata2$std_res <- mydata2$error / rmse
> mydata2$abs_std_res <- abs(mydata2$error / rmse)
> head(mydata2)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new predicted_value   error per_abs_error   sqerror    std_res abs_std_res
1        201   28  59      196     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
2        202   28  58       98     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
3        203   59  66      166     No       No      No          0            0           0          41.227  17.773     0.30123729 315.87953  2.0760798   2.0760798
4        204   65  67      163     No       No     Yes          0            0           1          60.834   4.166     0.06409231  17.35556  0.4866341   0.4866341
5        205   64  78      120     No       No     Yes          0            0           1          60.834   3.166     0.04946875  10.02356  0.3698232   0.3698232
6        206   59  57      152     No      Yes      No          0            1           0          54.288   4.712     0.07986441  22.20294  0.5504129   0.5504129
> mydata2_new <- subset(mydata2,abs_std_res < 1.96)
> head(mydata2_new)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new predicted_value   error per_abs_error   sqerror    std_res abs_std_res
1        201   28  59      196     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
2        202   28  58       98     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
4        204   65  67      163     No       No     Yes          0            0           1          60.834   4.166     0.06409231  17.35556  0.4866341   0.4866341
5        205   64  78      120     No       No     Yes          0            0           1          60.834   3.166     0.04946875  10.02356  0.3698232   0.3698232
6        206   59  57      152     No      Yes      No          0            1           0          54.288   4.712     0.07986441  22.20294  0.5504129   0.5504129
7        207   45  58      155     No      Yes      No          0            1           0          54.288  -9.288     0.20640000  86.26694 -1.0849395   1.0849395
> mydata2_new$per_abs_error <- abs((mydata2_new$risk - mydata2_new$predicted_value)/ mydata2_new$risk)
> mape <- mean(mydata2_new$per_abs_error)*100
> mape
[1] 12.30045


> mydata2_new$smk_age <- mydata2_new$Smoker_new*mydata2_new$Age
> mydata2_new$dia_age <-mydata2_new$Diabetes_new*mydata2_new$Age
> mydata2_new$fam_his_age <- mydata2_new$Fam_his_new*mydata2_new$Age
> mydata2_new$smk_pre <- mydata2_new$Smoker_new*mydata2_new$Pressure
> mydata2_new$diab_pre <-mydata2_new$Diabetes_new*mydata2_new$Pressure
> mydata2_new$famhis_pre <- mydata2_new$Fam_his_new*mydata2_new$Pressure
> head(mydata2_new)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new predicted_value   error per_abs_error   sqerror    std_res abs_std_res
1        201   28  59      196     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
2        202   28  58       98     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
4        204   65  67      163     No       No     Yes          0            0           1          60.834   4.166     0.06409231  17.35556  0.4866341   0.4866341
5        205   64  78      120     No       No     Yes          0            0           1          60.834   3.166     0.04946875  10.02356  0.3698232   0.3698232
6        206   59  57      152     No      Yes      No          0            1           0          54.288   4.712     0.07986441  22.20294  0.5504129   0.5504129
7        207   45  58      155     No      Yes      No          0            1           0          54.288  -9.288     0.20640000  86.26694 -1.0849395   1.0849395
  dia_age fam_his_age smk_pre diab_pre famhis_pre smk_age
1       0          0       0        0          0       0
2       0          0       0        0          0       0
4       0         67       0        0        163       0
5       0         78       0        0        120       0
6      57          0       0      152          0       0
7      58          0       0      155          0       0
```

```
> mymodel2_new=lm(risk ~.,data=mydata2_new)
> summary(mymodel2_new)

Call:
lm(formula = risk ~ ., data = mydata2_new)

Residuals:
       Min          1Q      Median          3Q         Max
-9.420e-14 -6.659e-15 -2.110e-16  5.875e-15  1.113e-13

Coefficients: (5 not defined because of singularities)
                 Estimate Std. Error    t value Pr(>|t|)
(Intercept)     4.123e+01  4.061e-13  1.015e+14  < 2e-16 ***
patient_id     -3.139e-16  5.585e-16 -5.620e-01  0.58006
Age             2.560e-15  5.665e-15  4.520e-01  0.65604
Pressure        9.361e-16  3.280e-16  2.854e+00  0.00949 **
SmokerYes       2.366e+01  3.113e-13  7.601e+13  < 2e-16 ***
DiabetesYes     1.306e+01  2.387e-13  5.471e+13  < 2e-16 ***
Fam_hisYes      1.961e+01  6.680e-13  2.935e+13  < 2e-16 ***
Smoker_new             NA         NA         NA       NA
Diabetes_new           NA         NA         NA       NA
Fam_his_new            NA         NA         NA       NA
predicted_value        NA         NA         NA       NA
error           1.000e+00  2.117e-15  4.723e+14  < 2e-16 ***
per_abs_error   2.858e-13  3.038e-13  9.410e-01  0.35759
sqerror         1.055e-16  1.053e-15  1.000e-01  0.92118
std_res                NA         NA         NA       NA
abs_std_res    -4.969e-14  9.286e-14 -5.350e-01  0.59822
dia_age         5.932e-17  2.029e-15  2.900e-02  0.97695
fam_his_age     3.315e-15  5.473e-15  6.060e-01  0.55124
smk_pre        -8.944e-16  4.586e-16 -1.950e+00  0.06465 .
diab_pre       -2.496e-16  7.700e-16 -3.240e-01  0.74903
famhis_pre      7.046e-16  1.824e-15  3.860e-01  0.70319
smk_age        -3.227e-15  4.607e-15 -7.000e-01  0.49132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.445e-14 on 21 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.414e+29 on 16 and 21 DF,  p-value: < 2.2e-16

Warning message:
In summary.lm(mymodel2_new) :
  essentially perfect fit: summary may be unreliable
> attach(mydata2_new)
The following objects are masked from mydata2:

    Age, Diabetes, Diabetes_new, Fam_his, Fam_his_new, patient_id, Pressure, risk, Smoker, Smoker_new


> mydata2_new$predictedvalue_new=41.23+(9.361e-16*Pressure)+(23.66*Smoker_new)+(13.06*Diabetes_new)+(19.61*Fam_his_new)+(-8.944e-16*smk_pre)+(error)
> head(mydata2_new)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new predicted_value   error per_abs_error   sqerror    std_res abs_std_res
1        201   28  59      196     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
2        202   28  58       98     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
4        204   65  67      163     No       No     Yes          0            0           1          60.834   4.166     0.06409231  17.35556  0.4866341   0.4866341
5        205   64  78      120     No       No     Yes          0            0           1          60.834   3.166     0.04946875  10.02356  0.3698232   0.3698232
6        206   59  57      152     No      Yes      No          0            1           0          54.288   4.712     0.07986441  22.20294  0.5504129   0.5504129
7        207   45  58      155     No      Yes      No          0            1           0          54.288  -9.288     0.20640000  86.26694 -1.0849395   1.0849395
  dia_age fam_his_age smk_pre diab_pre famhis_pre smk_age predictedvalue_new
1       0           0       0        0          0       0             28.003
2       0           0       0        0          0       0             28.003
4       0          67       0        0        163       0             65.006
5       0          78       0        0        120       0             64.006
6      57           0       0      152          0       0             59.002
7      58           0       0      155          0       0             45.002
> mydata2_new$per_abs_error_new <- abs((mydata2_new$risk - mydata2_new$predictedvalue_new)/ mydata2_new$risk)
> head(mydata2_new)
  patient_id risk Age Pressure Smoker Diabetes Fam_his Smoker_new Diabetes_new Fam_his_new predicted_value   error per_abs_error   sqerror    std_res abs_std_res
1        201   28  59      196     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
2        202   28  58       98     No       No      No          0            0           0          41.227 -13.227     0.47239286 174.95353 -1.5450575   1.5450575
4        204   65  67      163     No       No     Yes          0            0           1          60.834   4.166     0.06409231  17.35556  0.4866341   0.4866341
5        205   64  78      120     No       No     Yes          0            0           1          60.834   3.166     0.04946875  10.02356  0.3698232   0.3698232
6        206   59  57      152     No      Yes      No          0            1           0          54.288   4.712     0.07986441  22.20294  0.5504129   0.5504129
7        207   45  58      155     No      Yes      No          0            1           0          54.288  -9.288     0.20640000  86.26694 -1.0849395   1.0849395
  dia_age fam_his_age smk_pre diab_pre famhis_pre smk_age predictedvalue_new per_abs_error_new
1       0           0       0        0          0       0             28.003      1.071429e-04
2       0           0       0        0          0       0             28.003      1.071429e-04
4       0          67       0        0        163       0             65.006      9.230769e-05
5       0          78       0        0        120       0             64.006      9.375000e-05
6      57           0       0      152          0       0             59.002      3.389831e-05
7      58           0       0      155          0       0             45.002      4.444444e-05
> mape_new <- mean(mydata2_new$per_abs_error_new)*100
> mape_new
[1] 0.0050225
```

INTERPRETATION

MODEL SUMMARY:

Model	R	R Square	Adjusted R Square	F value	Significant value (p value)	Results
1	0.8742	0.7643	0.7296	22.05	8.605e-10	Significant

Significance at 1% level

COEFFICIENT TABLE:

	Estimate	Std. Error	t value	Pr (>|t|)	Result
Intercept	30.28452	18.36081	1.649	0.108	Insignificant
Age	0.14159	0.22693	0.624	0.537	Insignificant
Pressure	0.01584	0.05058	0.313	0.756	Insignificant
Smoker_new	21.44273	4.59084	4.671	4.57e-05	Significant at 0.1%
Diabetes_new	13.00869	2.93768	4.428	9.34e-05	Significant at 0.1%
Fam_his_new	18.50377	3.56991	5.183	9.93e-06	Significant at 0.1%

Here the dependent variable is 'risk' and the independent variables are age, pressure, smoker_new, diabetes_new and fam_his_new. First we converted the categorical variables into numerical variables. The variables age and pressure are insignificant. A smoker has 21.44273 times risk than a non-smoker. A diabetic person has 13.00869 times risk than a non-diabetic person. Similarly, a person with a family history of the same problem has 18.50377 times risk.

A new variable' predicted_value' is created in the table and the values for the dependent variable 'risk' is predicted using R code. The predicted values for the first 6 patients are as follows:

	Estimate	Std. Error	t value	Pr (>|t|)	Result
Intercept	30.28452	18.36081	1.649	0.108	Insignificant
Age	0.14159	0.22693	0.624	0.537	Insignificant
Pressure	0.01584	0.05058	0.313	0.756	Insignificant
Smoker_new	21.44273	4.59084	4.671	4.57e-05	Significant at 0.1%
Diabetes_new	13.00869	2.93768	4.428	9.34e-05	Significant at 0.1%
Fam_his_new	18.50377	3.56991	5.183	9.93e-06	Significant at 0.1%

We found out mean absolute error and mean absolute percentage error. The values ae as follows MAE= 7.0625 , MAPE= 13.19161

The outliers are detected and they are removed from the model. Then mean absolute percentage error is calculated. It was reduced to 12.30045

Interactions between the variables are calculated and thus a new model is built.

MODEL SUMMARY

Model	R	R Square	Adjusted R Square	F value	Significant value (p value)	Results
2	1	1	1	6.414e+29	<2.2e-16	Significant

Significant at 1% level

COEFFICIENT TABLE

	Estimate	Std. Error	t value	Pr (>|t|)	Result
Intercept	4.123e+01	4.061e-13	1.015e+14	<2e-16	Significant at 0.1%
Age	2.560e-15	5.665e-15	4.520e-01	0.65604	Insignificant
Pressure	9.361e-16	3.280e-16	2.854e+00	0.00949	Significant at 1%
Smoker_new	2.366e+01	3.113e-13	7.601e+13	<2e-16	Significant at 0.1%
Diabetes_new	1.306e+01	2.387e-13	5.471e+13	<2e-16	Significant at 0.1%
Fam_his_new	1.961e+01	6.680e-13	2.935e+13	<2e-16	Significant at 0.1%

smk_age	-3.227e-15	4.607e-15	-7.000e-01	0.49132	Insignificant
dia_age	5.932e-17	2.029e-15	2.900e-02	0.97695	Insignificant
famhis_age	3.315e-15	5.473e-15	6.060e-01	0.55124	Insignificant
smk_pre	-8.944e-16	4.586e-16	-1.950e+00	0.06465	Significant at 10%
diab_pre	-2.496e-16	7.700e-16	-3.240e-01	0.74903	Insignificant
famhis_pre	7.046e-16	1.824e-15	3.860e-01	0.70319	Insignificant
error	1.000e+00	2.117e-15	4.723e+14	<2e-16	Significant at 0.1%

Here the variable Pressure, Smoker_new, Diabetes_new, Fam_his_new and smk_pre are significant. New predicted values are calculated using the new model. The mean absolute percentage error for the new model is 0.0050225