Lund University
School of Economics and Management

Hita_Totaro_A1

Xhesina Hita, Paolo Totaro

2023-09-20

Problem 1: Linear Regression

Own model

In this first part was required to choose 3 of all the numerical independent variables. We start with an exploratory analysis of the data, we take into account the correlation between the numerical independent variables and our dependent variable Sales (unit sales at each location). We consider the values of the correlation found in the correlation matrix, correlation plot and real-life context. Then we select the variables which are more correlated with the dependent variable.

Note: In this preliminary analysis we make no use of any model or any kind of tests.

Taking a look at the correlation plot (Figure 1) we expect a linear relationship between Sales and Price (price company charges for car seats at each site), a quite strong negative correlation, since as we might expect in a real-life situation, as the price increases the unit sales decrease. When checking the value in the correlation matrix (Table 1) we notice indeed a value of -0.4546.

The other variable that seems to be significantly correlated with Sales is Advertising (local advertising budget for company), as we can see from the plot, we await a positive moderate linear relation, therefore the higher the budget in advertising, the higher the number of sales of car seats; the value found in the correlation matrix is 0.2846. We believe that a higher budget for advertising increases the awareness of the customers regarding safety system for children and therefore it will make them more interested in the product.

When considering the correlation between Age (average age of the local population) and Sales we don't notice from the correlation plot any significant trend, perhaps we catch a glimpse of a linear decrease in sales as the age increases. By taking a look at the correlation matrix with the actual values, we see a moderate negative relation between the 2 variables of -0.2394, this is a reasonable result since we expect that an older population corresponds to a lower number of children and therefore a lower number of sales of car seats. On the other hand, of course, a younger population will have more children and therefore a higher need of car seats.

We don't expect any multicollinearity since, by looking at the correlation matrix (Table 1), the 3 variables aren't correlated one another.

We left out of the model 4 variables since they don't seem to be correlated with the dependent variable Sales; in the correlation plot (Figure 1) we don't see any particular trend as each scatter plot seems to be very random.

Table 1: Correlation matrix

	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
Sales	1.0000000	0.0444125	0.1264934	0.2846205	0.0380768	-0.4546051	-0.2393756	-0.0396118
CompPrice	0.0444125	1.0000000	-0.0784934	-0.0220778	-0.0969067	0.5938969	-0.1538232	0.0643483
Income	0.1264934	-0.0784934	1.0000000	0.0768610	-0.0164522	-0.0586388	0.0358481	-0.0928130
Advertising	0.2846205	-0.0220778	0.0768610	1.0000000	0.2664988	0.0398910	-0.0026030	-0.0255342
Population	0.0380768	-0.0969067	-0.0164522	0.2664988	1.0000000	-0.0118726	-0.0178373	-0.1488199
Price	-0.4546051	0.5938969	-0.0586388	0.0398910	-0.0118726	1.0000000	-0.1163017	-0.0048697
Age	-0.2393756	-0.1538232	0.0358481	-0.0026030	-0.0178373	-0.1163017	1.0000000	0.0482994
Education	-0.0396118	0.0643483	-0.0928130	-0.0255342	-0.1488199	-0.0048697	0.0482994	1.0000000

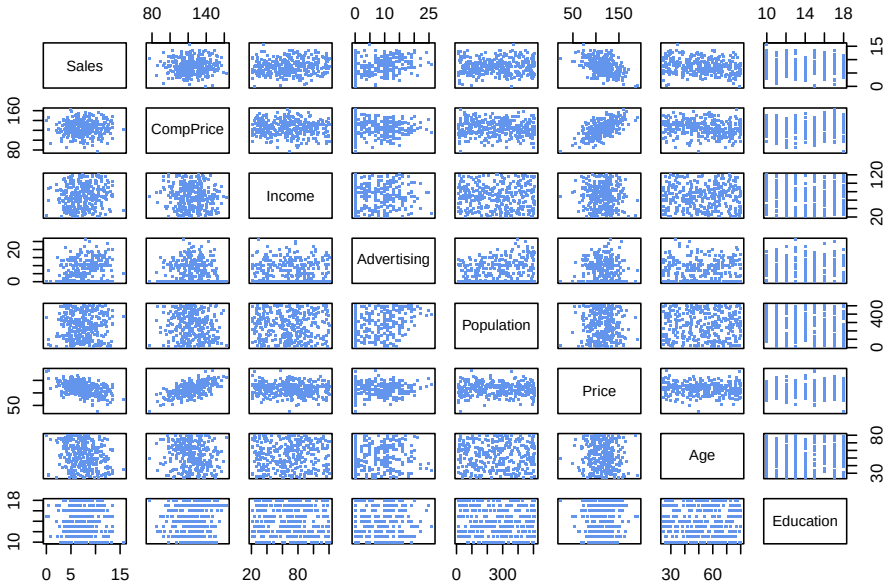


Figure 1: Correlation plot

Table 2: Reduced model coefficients (x)

	x
(Intercept)	15.9451440
Price	-0.0578174
Advertising	0.1290287
Age	-0.0503991

Table 3: Confidence intervals of reduced model coefficients

	2.5 %	97.5 %
(Intercept)	14.4379171	17.4523709
Price	-0.0679021	-0.0477326
Advertising	0.0921526	0.1659047
Age	-0.0652296	-0.0355685

Table 2 shows the coefficients of the variables, it can be observed that advertising has a greater effect on Sales than the other two variables.

Table 3 supplies us with the values of the confidence intervals of the reduced model coefficients at 95% level, which means that there is approximately 95% chance that the given interval will contain the true values of the model coefficients. Since the confidence intervals of the three predictors don't contain the 0 value, we expect that the p-values of the t-tests will be significant.

When fitting the model, the value of the F-statistic (66.06 with 3 and 316 degree of freedom) is pretty high, therefore the p-value is extremely low (2.2 e-16), meaning that at least one of the 3 coefficients of the variables is significantly different from 0 at a confidence level greater than 95%. When considering the t-values of each variable taken individually we also get pretty high results in term of absolute value, therefore the p-values are very low as well, meaning that each predictor has a significant effect on our dependent variable Sales again at confidence level of 95%.

```
Call:
lm(formula = Sales ~ Price + Advertising + Age, data = CarseatsS)

Residuals:
   Min     1Q Median     3Q    Max
-5.714 -1.502 -0.100  1.471  5.816

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.945144   0.766062  20.814  < 2e-16 ***
Price       -0.057817   0.005126 -11.280  < 2e-16 ***
Advertising  0.129029   0.018743   6.884 3.14e-11 ***
Age         -0.050399   0.007538  -6.686 1.04e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.17 on 316 degrees of freedom
Multiple R-squared:  0.3854,    Adjusted R-squared:  0.3796
F-statistic: 66.06 on 3 and 316 DF,  p-value: < 2.2e-16
```

We compute the Variance Inflation Factor for each predictor, values close to 1 indicate the absence of collinearity. For our model we get values very close to 1, therefore the hypothesis of multicollinearity can be dropped as we expected from the Correlation plot.

	Price	Advertising	Age
VIF	1.015325	1.001598	1.013716

We now proceed with the check of the 5 assumptions for the linear regression:

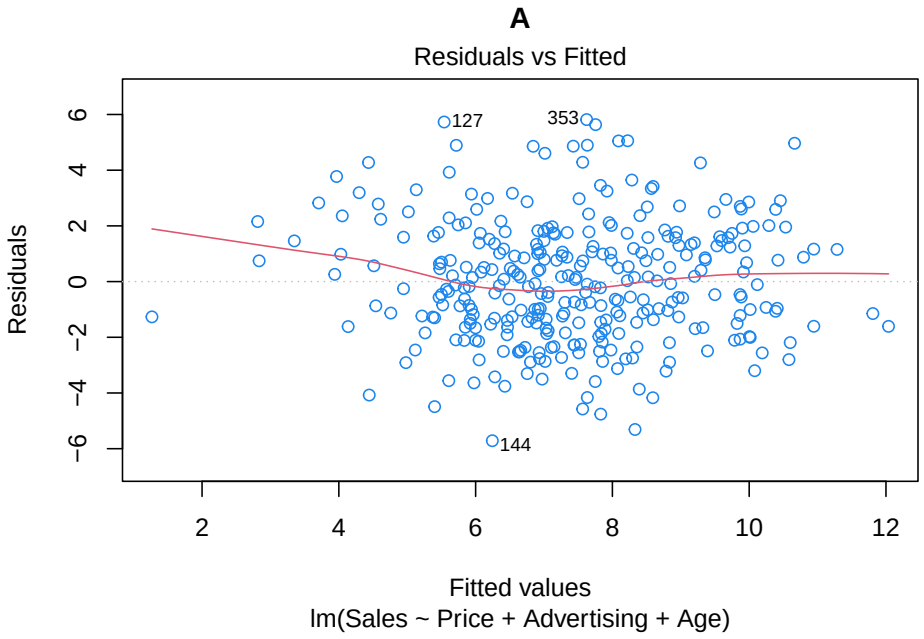
1) We build the plot of Residuals versus Fitted Values (A) in order to check the linear relationship between predictors and the dependent variable. As we can see from the generated plot, there seems to be a decreasing trend at the beginning which anyway stabilises around the 0 for the majority of the plot. Therefore since most of the residuals lie around the 0 line we consider appropriate the linearity assumption.

2) In order to check the homoscedasticity assumption that the disturbance terms have all the same variance and are not correlated to one another we build the Scale-Location plot (B) and check the pattern of the fitted line. In our case we get a pretty straight line since our points are all equally spread around the value 1, therefore there is no violation of the homoscedasticity assumption.

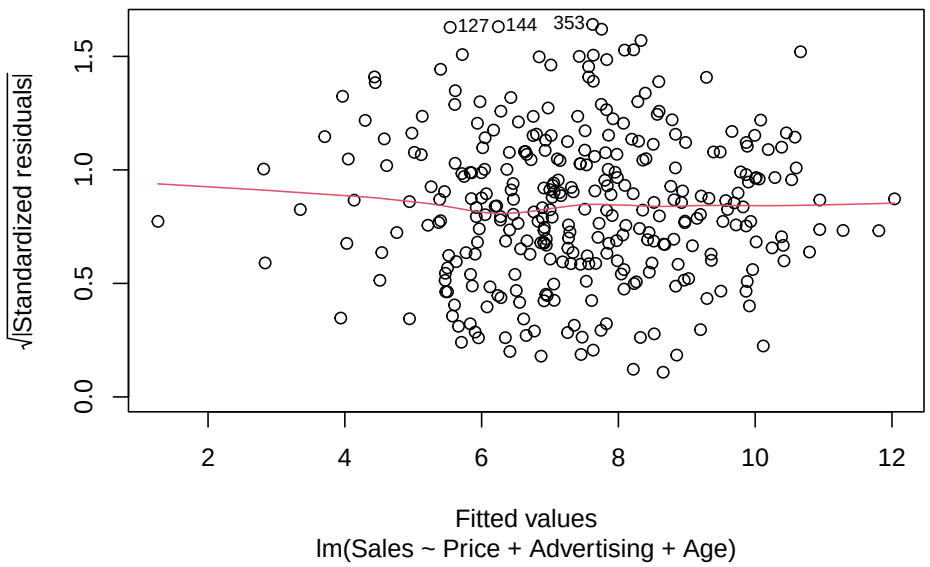
3) The normality assumption implies that the residuals of our model are normally distributed, in order to check this, we draw the QQ-plot (C) and check whether the quantiles of a standardized residuals approximately correspond to the quantiles of a standard normal distribution (the points in the QQ-plot lie inside the confidence interval). This is the case of our model, therefore, the normality assumption seems reasonable. In order to be more confident of our statement we can check the histogram of the residuals (D) and see whether they follow a normal distribution. In our case they do and therefore the assumption does not seem to be violated.

4) As we can see from the histogram (D) the expected value of the disturbance term is 0 hence the assumption on the exogeneity of the independent variables seems reasonable.

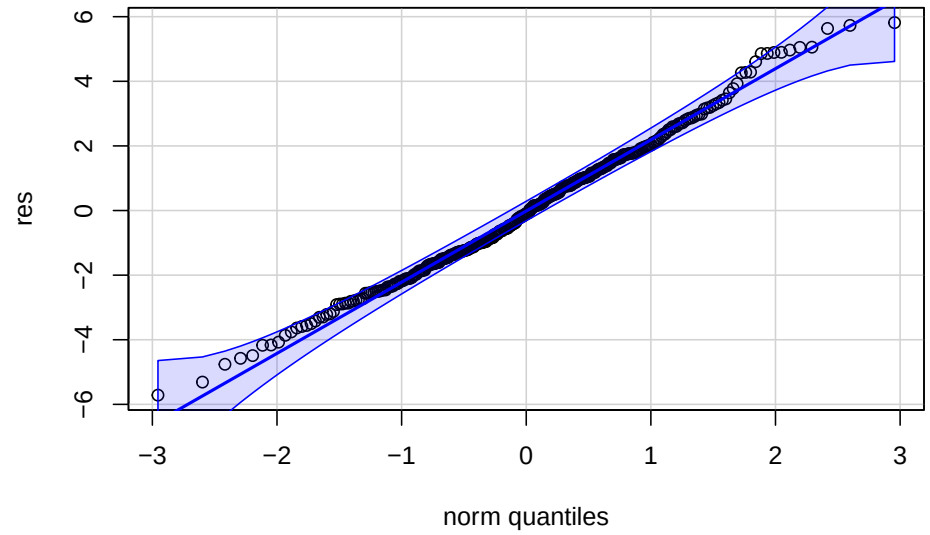
5) The last assumption to verify is that there is no exact linear relationship among any of the independent variable. In order to check this we inspect the correlation matrix (Table 1) of the predictors of our model and see if there is any linear correlation between Age, Advertising and Price. There isn't any, therefore, the assumption can be considered appropriate.

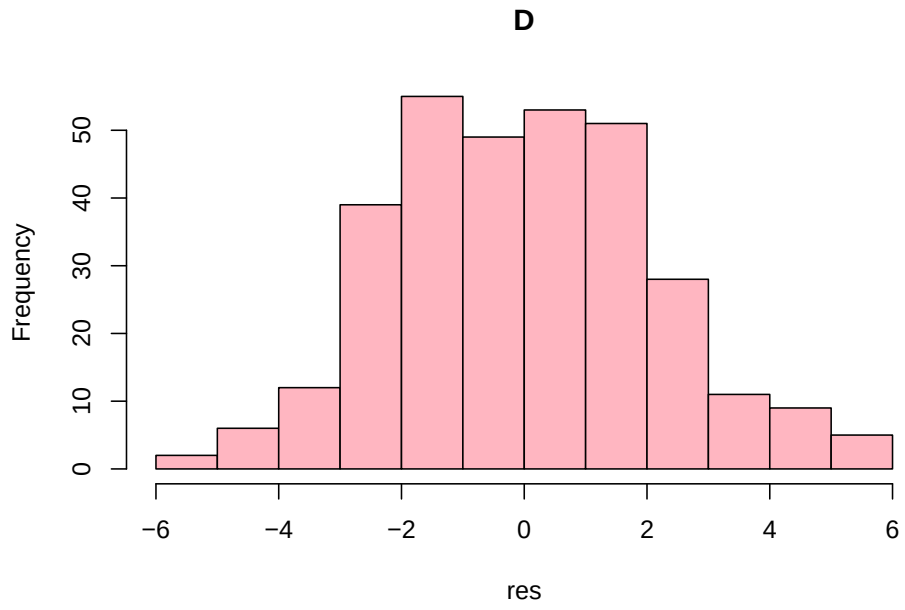


B


Scale−Location
127
144
353
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
4
6
8
10
12
Fitted values
lm(Sales ~ Price + Advertising + Age)

C


res
−6
−4
−2
0
2
4
6
−3
−2
−1
0
1
2
3
norm quantiles



All possible variables

First of all, we fit the full model with all the 7 numerical predictors. We consider the F-statistic and notice how big it is, with a very small p-value, this means that there is at least one coefficient significantly different from 0 at 95% level. Indeed, if we observe the p-values for the t-tests there are 5 significant beta coefficients, the only predictors which seem not to have a significant impact on Sales are Education (education level at each location) and Population (population size in a region). If we work with a significance level of 99%, the predictor Income is not significant anymore.

We notice that even if we include all the numerical variables in the model, the ones we chose for the reduced one (Price, Advertising and Age) remain highly significant. Advertising remains the variable that has the greater effect on Sales.

```
Call:
lm(formula = Sales ~ ., data = CarseatsS)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0922 -1.3259 -0.1689  1.0979  4.9641

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.2313685  1.2297243   6.694 1.01e-10 ***
CompPrice    0.0893067  0.0089328   9.998  < 2e-16 ***
Income       0.0099479  0.0039183   2.539   0.0116 *
Advertising  0.1368815  0.0169645   8.069 1.55e-14 ***
```

```
Population  -0.0003722  0.0007424  -0.501   0.6165
Price       -0.0898855  0.0055165 -16.294  < 2e-16 ***
Age         -0.0435162  0.0065981  -6.595 1.82e-10 ***
Education   -0.0497932  0.0406871  -1.224   0.2219
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.884 on 312 degrees of freedom
Multiple R-squared:  0.5424,    Adjusted R-squared:  0.5322
F-statistic: 52.84 on 7 and 312 DF,  p-value: < 2.2e-16
```

Since we are now considering the full model, we also have to look for multicollinearity which is not a good factor for our model. In this case we have multicollinearity, by taking a look at the correlation plot (Figure 1) of all the independent variables we notice that the only noticeable correlation is the one among Price and CompPrice (price charged by competitor at each location). The value of their correlation in 0.59 which is very high in comparison with the correlations among the other predictors, therefore we expect there is multicollinearity between these 2 variables.

Another way to check for multicollinearity is buy exploiting Pearson correlation. From Table A we observe through the p-values that Advertising, Price, Age and Income are highly correlated with the dependent variable. Then, if we consider the correlation among the independent variables, we see that Population is highly correlated with Advertising, CompPrice with Price, Age with CompPrice, Education with Population and Price with Age at significance level of 95%. Therefore, there is multicollinearity in the full model.

Table A: p-values from Pearson Correlation

	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
P.Sales	NA	0.429	0.024	0.000	0.497	0.000	0.000	0.480
P.CompPrice	0.429	NA	0.161	0.694	0.083	0.000	0.006	0.251
P.Income	0.024	0.161	NA	0.170	0.769	0.296	0.523	0.097
P.Advertising	0.000	0.694	0.170	NA	0.000	0.477	0.963	0.649
P.Population	0.497	0.083	0.769	0.000	NA	0.832	0.751	0.008
P.Price	0.000	0.000	0.296	0.477	0.832	NA	0.038	0.931
P.Age	0.000	0.006	0.523	0.963	0.751	0.038	NA	0.389
P.Education	0.480	0.251	0.097	0.649	0.008	0.931	0.389	NA

We now consider the Variance Inflation Factor in order to check our assumptions of multicollinearity. In Table B we see indeed that CompPrice and Price have the highest VIF and therefore we expect multicollinearity in the model.

Table B: Variance Inflation Factors (VIF) of the dependent variables

	CompPrice	Income	Advertising	Population	Price	Age	Education
VIF	1.5963	1.0245	1.0882	1.1142	1.5596	1.03	1.0404

We then check the assumptions of linear regression in the full model like we did for the reduced one, turns out the plots are quite similar and therefore the assumptions are fulfilled like in the reduced model.

After fitting the full model, we can compare it with the reduced one. Both models fulfil the assumptions of linear regression, except for the full model where multicollinearity is present. In the full model we have too many predictors since some of them do not

have any significant effect on the dependent variable Sales. When taking a look at the coefficient of Determination (R squared) we notice that in the full model it is 0.542 whereas in the reduced it is 0.385. This is due to the fact that when adding predictors to a model the value of R squared increases. If we consider the adjusted R squared that doesn't take into account the number of independent variables we have an increase of R squared anyway, this is due to the fact that part of the variability that was hidden in the disturbance term is now explained by some of the newly included predictors.

We also check the value of the Residual Standard Error to get an estimate of the standard deviation of epsilon (error) from the average of the dependent variable Sales (Table C). The RSE for the reduced model is 2.17 and the RSE of the full model is 1.884. We now compute the percentage error by dividing each of these results by the average Sales (7.432); we then obtain the error percentages: 29.20% for the reduced model and 25.35% for the full one.

Table C: Residual Standard Error of the two models (Reduced and Full)

Average_Sales	Percentage_Error_REDUCED	Percentage_Error_FULL
7.431906	29.19843	25.35016

Problem 2: Classification

KNN

We now perform a K-nearest neighbor classification analysis. Our dependent variable is now High which assume value 1 if the number of unit sales is greater than 8 thousands and 0 otherwise. We train the model on the numerical predictors (standardized since they are in different scales) in the train data set in order to classify our data. The train data set is obtained considering part of the observations from the full data set, in our case 320 out of 400.

We take odds values of k ranging from 1 to 13 (7 values), we only take the odds values to not get randomness in classification due to even values of k. From this analysis we get that the best classification comes from k=9 as we can see both in Table D and Figure E.

Table D: Proportion of correct classifications for different values of K

propor_1	propor_3	propor_5	propor_7	propor_9	propor_11	propor_13
0.6625	0.6375	0.7	0.6875	0.725	0.675	0.6375

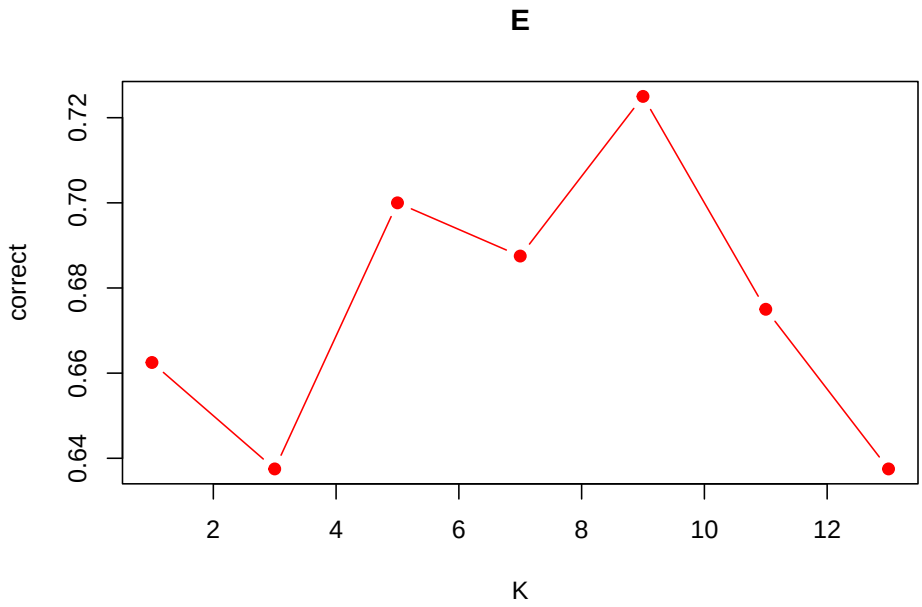


Figure E: Proportion of Correct Classification with KNN

We also performed KNN using Leave One Out Cross Validation in order to check the stability of our results using k values from 1 to 13. After setting a seed, we get that the best classification comes from k=11 as we can see from Table E and Figure F.

Table E: Proportion of correctly classified values using Cross Validation

K	cv_propor
1	0.655
2	0.670
3	0.670
4	0.665
5	0.677
6	0.705
7	0.688
8	0.695
9	0.710
10	0.710
11	0.718
12	0.703
13	0.713

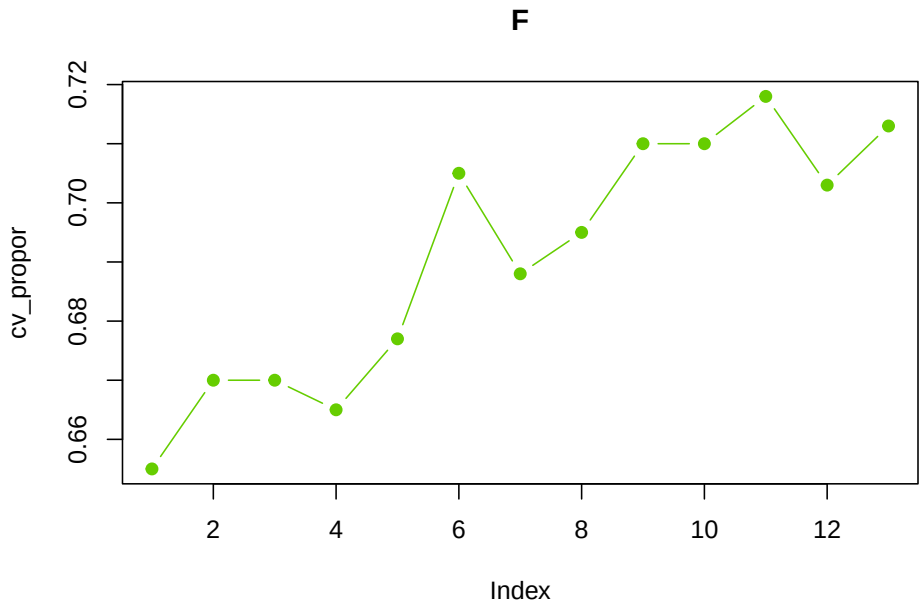


Figure F: Proportion of Correct Classification with Cross Validation

On average the two classification approaches are very similar, we see a slightly higher proportion value in the first approach using k=9.

Logistic Regression

We now perform Multiple Logistic Regression on our data set. We carry out three classifications according to 3 different cut-points(0.5, 0.3 and 0.8) using High as dependent variable and the other numerical variables as predictors. We then obtain the values of Sensitivity and Specificity for each cut-point:

- **0.5 cut-point**: Sensitivty=84.375% and Specificity=75% . The values we obtained correspond respectively to the *true positive rate* and *true negatvie rate*.
- **0.3 cut-point**: Sensitivity=68.889% and Specificity=77.143%
- **0.8 cutpoint**: Sensitivity=94.444% ans Specificity=64.516%

Cutpoint 0.5

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  80
```

```
             | High.f[-train]
      pred.c |       No  |      Yes  | Row Total |
-------------|-----------|-----------|-----------|
          No |       36  |       12  |       48  |
             |   75.000% |   25.000% |   60.000% |
-------------|-----------|-----------|-----------|
         Yes |        5  |       27  |       32  |
             |   15.625% |   84.375% |   40.000% |
-------------|-----------|-----------|-----------|
Column Total |       41  |       39  |       80  |
-------------|-----------|-----------|-----------|
```

Cutpoint 0.3

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  80

             | High.f[-train]
     pred.c3 |       No  |      Yes  | Row Total |
-------------|-----------|-----------|-----------|
          No |       27  |        8  |       35  |
             |   77.143% |   22.857% |   43.750% |
-------------|-----------|-----------|-----------|
         Yes |       14  |       31  |       45  |
             |   31.111% |   68.889% |   56.250% |
-------------|-----------|-----------|-----------|
Column Total |       41  |       39  |       80  |
-------------|-----------|-----------|-----------|
```

Cutpoint 0.8

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  80
```

	High.f[-train]		
pred.c8	No	Yes	Row Total
No	40	22	62
	64.516%	35.484%	77.500%
Yes	1	17	18
	5.556%	94.444%	22.500%
Column Total	41	39	80

From Table E we get the proportion of correct classification for the three different cutpoints, we can observe that the best classification is obtained using 0.5 as cutpoint.

Table E: Proportion of correct classification for each cutpoint

propcorrect05	propcorrect03	propcorrect08
0.9125	0.725	0.7125

Figure 2 shows the plot of the ROC curve, we consider the area under the ROC curve (AUC) which represents the performance of the model, the bigger the area the better the performance. In this case the value of the AUC is 0.8380238

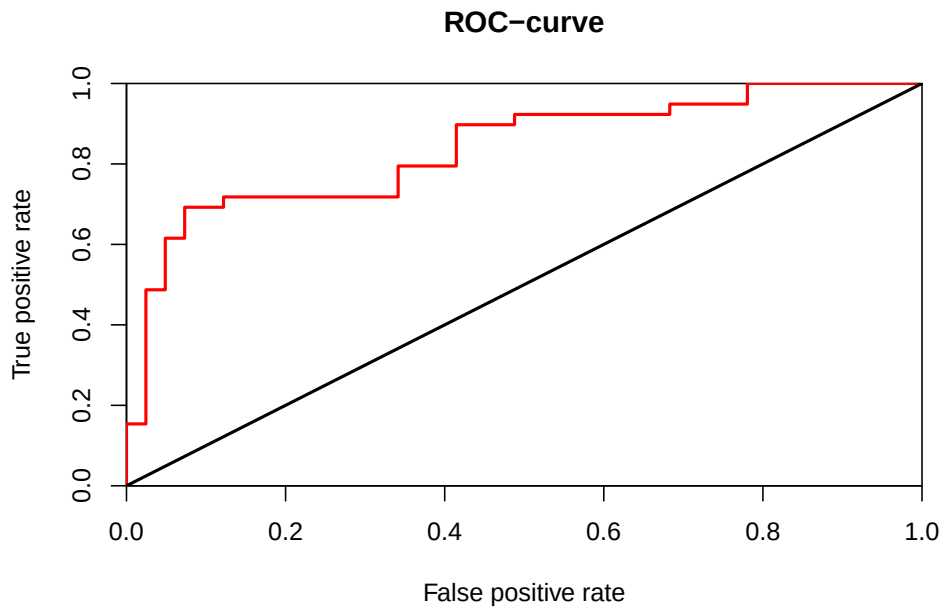


Figure 2: ROC-curve plot

In conclusion, we compare the different classification approaches and find the best one for our data. Taking into consideration KNN and KNN using Cross-Validation

we have already seen how there isn't a very high difference in proportion of correct classification, but if we had to choose we would use a KNN with k=9. We now compare the KNN with the Multiple Logistic Regression approach using the proportion of correct classification at cut-point 0.5 (since it is the best one among the three cutpoints we tried), the Multiple Linear Regression approach works extremely better than the KNN one, the proportions of correct classification are respectively 0.9125 and 0.725 . So for this data the best model seems to be a Multiple Logistic Regression with cut-point=0.5.