Lorenzo Pedrini student number: 20337113 date: 14/04/2024

ST8003 Linear Regression, 2023-24

Final Project

Part 1 – injury case

With this report we would like to assess the problematic of an unusual high rate of fatalities per burn injuries. Inefficient hospital care is considered a possible responsible for the problem. The investigation is necessary to identify risk factors in order to correctly allocate patients to emergency care when needed in order to avoid fatalities.

The dataset used for the analysis is "injury24", this dataset is publicly available as the file "injury24.xlsx"

The dataset is comprehensive of data from 1000 patients affected with burn injuries and it is composed of the following entries:

- fatality 1 = patient died, 0 = patient survived
- age age at admission in years
- male 1 = male, 0 = female
- race 1 = if white, 0 = non-white
- area burn surface area (% of the body)
- inhal 1 = if burn involved smoke inhalation, 0 = otherwise
- flame 1 = if burn involved flame, 0 = otherwise

The aim of the report is to fit a logistic regression to explain which predictors are statistically significant for predicting the outcome of a fatality.

Summary of the analysis and practical suggestions:

Best model:

$$\log\left(\frac{\pi}{1-\pi}\right) = -10.924 + 0.121 * \boldsymbol{age} + 0.273 * \boldsymbol{race} + 0.151 * \boldsymbol{area} + 6.801 * \boldsymbol{inhal} - 0.054 * \boldsymbol{race} * \boldsymbol{area} - 0.074 * \boldsymbol{age} * \boldsymbol{inhal} - 0.045 * \boldsymbol{area} * \boldsymbol{inhal}$$

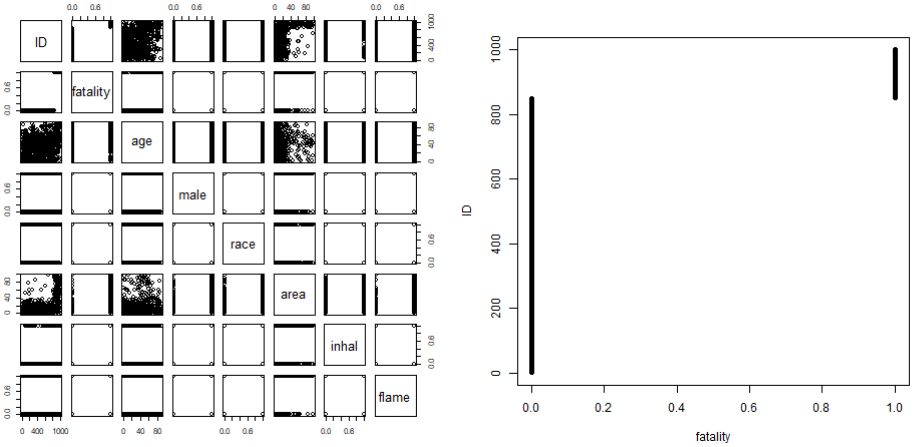
As the result of this statistical analysis, we can see the most relevant influence factor that explains the fatality of a burn accident is the inhalation of fumes. Immediate medical attention should be given for patients presenting this risk factor. Increasing age and increasing burn area are both correlated to an increased probability of fatality. Priority medical attention should be given to older patients and to patients with a high percentage area of the body covered by burns.

Exploratory analysis:

The scatterplot matrix highlights to interesting features of the dataset (Figure 1.1). The entries in the dataset show a clear order, the patients presenting a non-fatality event are entered first, while the last part of the database is descriptive only of the patients presenting a fatality event (Figure 1.2).

A randomized sampling will be compulsory before separating the training set to the test set.

Figure 1.1 Figure 1.2



Statical analysis:

The dataset has been randomly sampled to use 90% of the data for training the model and 10% for testing the model.

For fitting the model, **fatality** has been used as the response for the logistic regression.

A full model (**eMLR1**) was fit to assess the contribution of all the variables for explaining the variability of the response (Figure 1.3).

Figure 1.3

```
> summary(eMLR1)

Call:
glm(formula = fatality ~ . - ID, family = "binomial", data = train_edata1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.586939   0.722987 -10.494  < 2e-16 ***
age          0.081713   0.009005   9.074  < 2e-16 ***
male        -0.104421   0.326737  -0.320   0.7493
race        -0.717146   0.324816  -2.208   0.0273 *
area         0.087521   0.009432   9.279  < 2e-16 ***
inhal        1.631920   0.380915   4.284 1.83e-05 ***
flame        0.432481   0.368443   1.174   0.2405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 764.34  on 899  degrees of freedom
Residual deviance: 303.18  on 893  degrees of freedom
AIC: 317.18

Number of Fisher Scoring iterations: 7
```

From the full model we can see that the **intercept** and the variables **age**, **race**, **area** and **inhale** are significant. Presence of collinearity was excluded by Variance Inflator Factor analysis (VIF) (Figure 1.4).

Figure 1.4

```
> vif(eMLR1)
     age     male     race     area    inhal    flame
1.636355 1.075498 1.122036 1.445186 1.242767 1.123132
```

Step AIC analysis (both forward and backward) led to the same best model that features and AIC score of 314.7:

fatality ~ age + race+ inhal + area

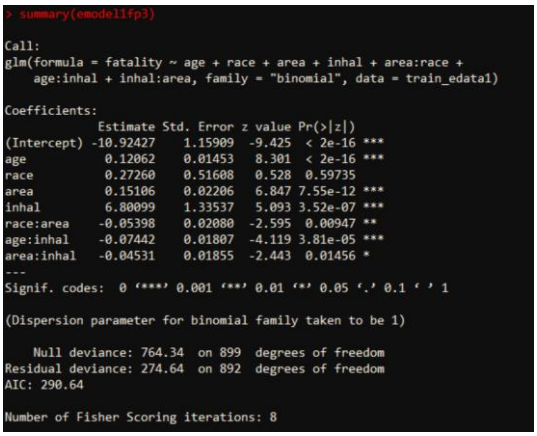
Model was fit as **emodel1fp1** and the results show that all the predictors are significant (Figure 1.5). The model presents a pseudo-R^2 of 60.1, all the variables are significant apart of **race** that is not significant. Before removing the variable **race** from the model, we proceeded with an analysis of the interaction between the variables.

Figure 1.5

```
> summary(emodel1fp2)

Call:
glm(formula = fatality ~ age + race + area + inhal + area:race,
    family = "binomial", data = train_edata1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.707840   0.793226 -10.978  < 2e-16 ***
age          0.087882   0.009249   9.502  < 2e-16 ***
race         0.540244   0.497092   1.087   0.2771
area         0.140033   0.020071   6.977 3.02e-12 ***
inhal        1.769820   0.370908   4.772 1.83e-06 ***
race:area   -0.065146   0.020251  -3.217   0.0013 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 764.34  on 899  degrees of freedom
Residual deviance: 292.34  on 894  degrees of freedom
AIC: 304.34

Number of Fisher Scoring iterations: 7
```

The interactions that were found significant are **area*race**, **age*inhal** and **inhal*area**. The model **emodel1fp3** was fit taking in account those interaction (Figure 1.6).

Figure 1.6



```
> summary(emodel1fp3)

Call:
glm(formula = fatality ~ age + race + area + inhal + area:race +
    age:inhal + inhal:area, family = "binomial", data = train_edata1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.92427    1.15909  -9.425  < 2e-16 ***
age           0.12062    0.01453   8.301  < 2e-16 ***
race          0.27260    0.51608   0.528  0.59735
area          0.15106    0.02206   6.847 7.55e-12 ***
inhal         6.80099    1.33537   5.093 3.52e-07 ***
race:area    -0.05398    0.02080  -2.595  0.00947 **
age:inhal    -0.07442    0.01807  -4.119 3.81e-05 ***
area:inhal   -0.04531    0.01855  -2.443  0.01456 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 764.34  on 899  degrees of freedom
Residual deviance: 274.64  on 892  degrees of freedom
AIC: 290.64

Number of Fisher Scoring iterations: 8
```

Model **emodel1fp3** presents an AIC score of 290.6, and a pseudo R^2 of 64.1. The variable **race** is still not significant, although the interaction between **race** and **area** is significant - residual plot shows values randomly sparce around zero and there is no evidence of the presence of high leverage points (Figure 1.7, Figure 1.8). A model without the variable **race** is fit to test if the removing it would give us a better model. The resulting model **emodel1fp6** is characterised by a higher AIC score (300.5) and a lower pseudo R^2 (62.3). Also an anova likelihood ratio test between the models pointed out the complex model **emodel1fp3** as the best model between the two (Figure 1.9).

Figure 1.7 Figure 1.8

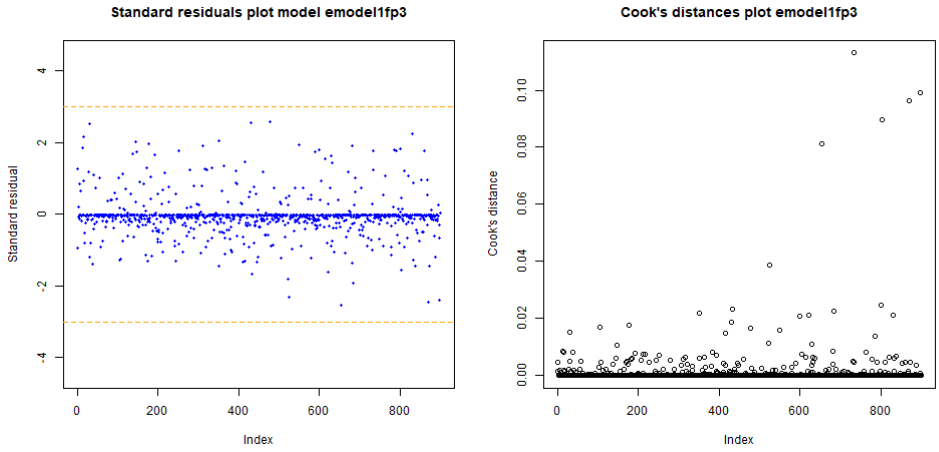


Figure 1.9

```
> anova(emodel1fp3,emodel1fp6,test="LRT")
Analysis of Deviance Table

Model 1: fatality ~ age + race + area + inhal + area:race + age:inhal +
    inhal:area
Model 2: fatality ~ age + area + inhal + age:inhal + inhal:area
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       892     274.64
2       894     288.48 -2  -13.839 0.0009884 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At the end of the analysis, we selected **emodel1fp3** as the best model for the study.

Best model:

$$\log\left(\frac{\pi}{1-\pi}\right) = -10.924 + 0.121 * \boldsymbol{age} + 0.273 * \boldsymbol{race} + 0.151 * \boldsymbol{area} + 6.801 * \boldsymbol{inhal} - 0.054 * \boldsymbol{race} * \boldsymbol{area} - 0.074 * \boldsymbol{age} * \boldsymbol{inhal} - 0.045 * \boldsymbol{area} * \boldsymbol{inhal}$$

Model interpretation:

The first considerations on the best model is that variables **flame** and **male** are not significant for predicting the fatality rate of burn injury patients. There are different possible causes of burn injury, they can be high temperature, electricity, friction, radiation and chemicals[1], possibly, flame relates just to a few of the entries on the dataset. The sex of a person doesn't seem to influence either the probability of fatality of a burn injury. **Age** seems to have a big influence on the response, the odds ratio at 20 years old is 11.25 (e^0.121*20), while the odds ratio for 40 years old is 126.47 (e^0.121*40). It implies that the odds are 10 times more in the case of the 40 years old, implying a very big implication of age on fatality. The binary variable **inhal** is significant and very influential over fatality outcome, the odds ratio for the presence of this factor makes the odds ratio to increase from 1 (e^0) to 898.75 (e^6.801). Interactions **inhal*age** and **inhal*area** are negative and they show that when patients have inhaled fumes, relation between fatality and age or area lowers, the latter variables become less influential over fatality when the **inhal** has a value of 1. Variable **race** is not too influential in the regression, the confidence interval overlaps the zero, although it is relevant the influence that race has to the are exposed to burns, looking at the interaction **race*area** white people seem to have a smaller burn injury area in relation to non-white people. Variable **area** has a similar influence than the variable **age** over the probability of a fatality event.

Model evaluation:

Predictions were made using the data of the test dataset formed of 100 observations – model **emodel1fp3** was used to predict the probability response. A cut-off value of 0.7 was chosen as a fatality threshold, any prediction over 0.7 was considered fatality, while any prediction lower than 0.7 was considered non fatality.

emodel1fp3 performance was evaluated using the confusion matrix (Figure X), leading to an accuracy of 91%. We consider this prediction accuracy satisfactory for the outcome of our analysis.

Part 2 – School case

This report is to answer the inquiry of the Prime Minister to review the allocation of the budget for the schools in Ireland, in particular we investigated how current budget allocation has been influenced student performance and suggested improvements to boost student achievements.

The dataset used for the analysis is the dataset "schools24" this dataset is publicly available as the file "schools24.xlsx".

The dataset is comprehensive of data from 170 schools in Ireland and it is composed of the following entries:

- spendingpp spending per student (€)
- spendspec spending per special need students (€)
- ipad students per ipad/computer
- special % of students with special needs
- free % of students eligible for free lunch
- stratio students per teacher ratio
- income school income per student from all sources (€000s)
- score1 standardised test score at 11
- score2 standardised test score at 16
- salary salary of teachers average (€000s)
- noteng % of students whose first language is not English

Two different models were fit, the first one related to the scores when students are 11 years old, while the second on is related to the score when students are 16. The fit of two different models will allow to understand if actual school budget distribution is effective or if the budget would need to be distributed differently.

Summary of the analysis and practical suggestions:

Best model:

score1 = 716.596 - 0.423 **special** - 0.478 **free** - 0.675 **stratio** - 0.55 **noteng** + 1.057 **income**

score2 = 676.930 + 1.799 **income** - 0.738 **free**

The result of the statistical analysis analysed two separate models, one analysing the scores of 11 years old students and the other one analysing the score of 16 year old students. Those results highlight that the most influent predictors that explains both the scores is the income of the school per students. It means that schools that have more money to invest per student have a high average value for both the scores. The amount of students eligible for free lunch is indicative of the presence of economically disadvantaged students in the school. This has a negative influence on the scores. Budget that is allocated for computers and teachers stipends doesn't have a statistical influence on the scores.

In summary it would be advisable to redistribute the budget, in the form of investing more of it into schools with a low income per student. This money can potentially be financed by a lowering in teacher's stipends and from a reduced investment on computers.

Exploratory analysis:

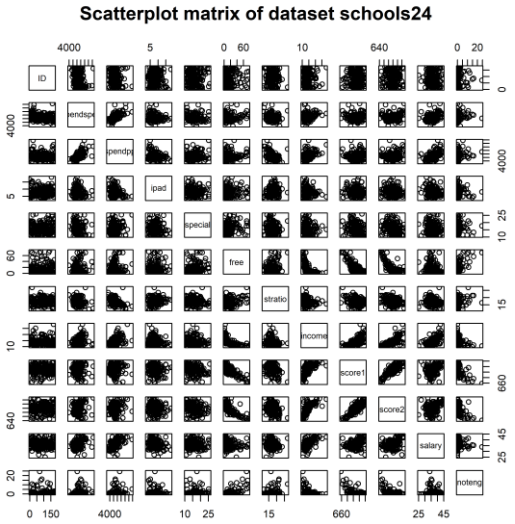
The scatterplot matrix shows a few evident linear relations between variables and responses (Figure 2.1).

School income per student clearly correlates with the percentage of students eligible for free lunch (Figure 2.2).

School income per student clearly relates with student score (Figure 2.3).

% of student eligible for a free lunch shows a clear linear relation with student score (Figure 2.4).

Figure 2.1 Figure 2.2



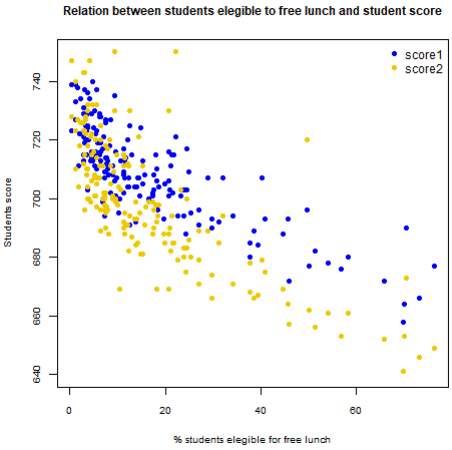
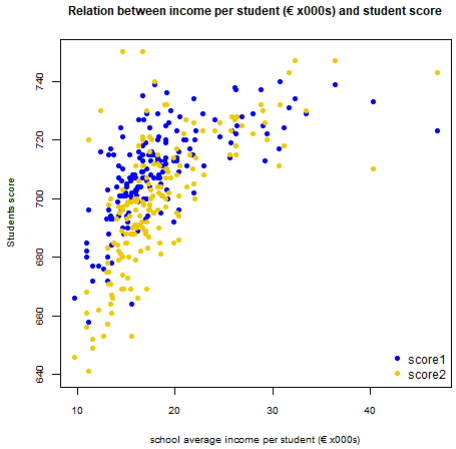
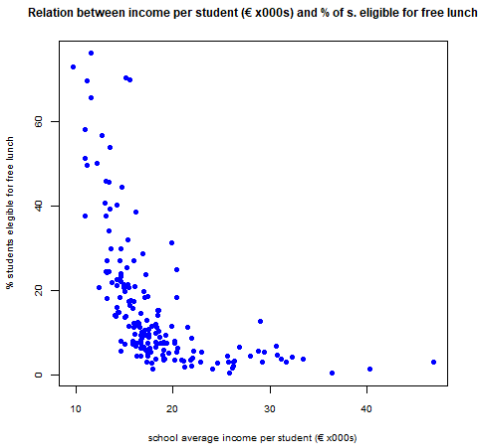


Figure 2.3 Figure 2.4





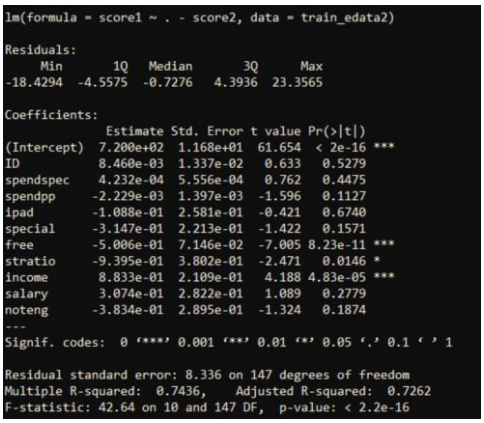
Statistical analysis:

The dataset has been randomly sampled to use 90% of the data for training the model and 10% for testing the model.

Model fit using **score1** as the response.

A full model **MLRs11** was fit to assess preliminary model performance when all the variables are included (Figure 2.5).

Figure 2.5

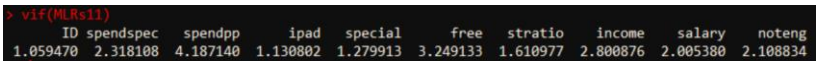


```
lm(formula = score1 ~ . - score2, data = train_edata2)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4294  -4.5575  -0.7276   4.3936  23.3565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.200e+02  1.168e+01  61.654  < 2e-16 ***
ID           8.460e-03  1.337e-02   0.633   0.5279
spendspec    4.232e-04  5.556e-04   0.762   0.4475
spendpp     -2.229e-03  1.397e-03  -1.596   0.1127
ipad        -1.088e-01  2.581e-01  -0.421   0.6740
special     -3.147e-01  2.213e-01  -1.422   0.1571
free        -5.006e-01  7.146e-02  -7.005 8.23e-11 ***
stratio     -9.395e-01  3.802e-01  -2.471   0.0146 *
income       8.833e-01  2.109e-01   4.188 4.83e-05 ***
salary       3.074e-01  2.822e-01   1.089   0.2779
noteng      -3.834e-01  2.895e-01  -1.324   0.1874
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.336 on 147 degrees of freedom
Multiple R-squared:  0.7436,    Adjusted R-squared:  0.7262
F-statistic: 42.64 on 10 and 147 DF,  p-value: < 2.2e-16
```

From the full model we can see that only the **intercept** and the variables **free**, **stratio** and **income** are significant.

Presence of collinearity was excluded with by Variance Inflator Factor analysis (VIF)(Figure 2.6).

Figure 2.6



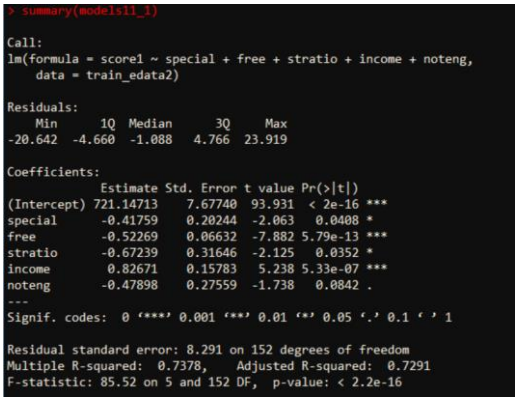
```
> vif(MLRs11)
      ID spendspec   spendpp      ipad   special      free   stratio    income    salary    noteng
1.059470  2.318108  4.187140  1.130802  1.279913  3.249133  1.610977  2.800876  2.005380  2.108834
```

Step AIC analysis (both forward and backward) outcome led to the same best fit model:

score1 ~ special + free + stratio + income + noteng

Model **Model11_1** was fit and the summary show that all the predictors are significant - the model explains 72.9% of the variability of the response **score1** (Figure 2.7).

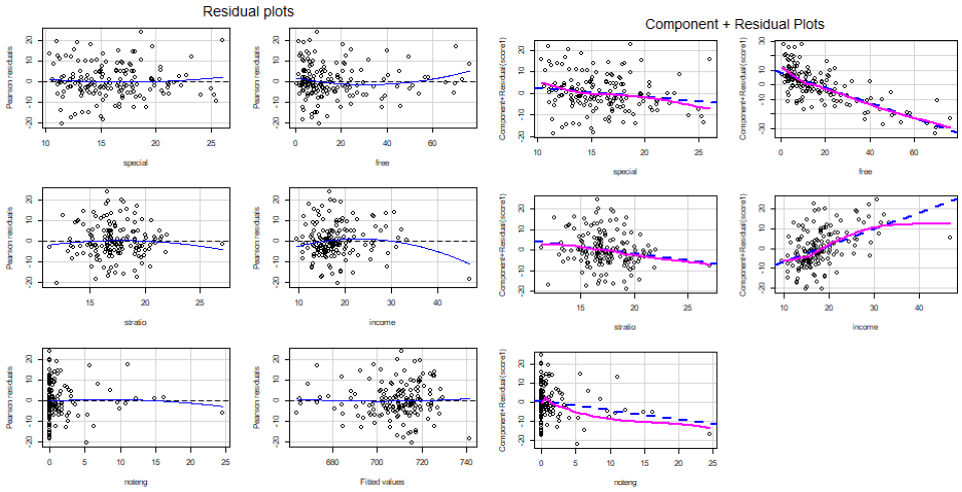
Figure 2.7



```
> summary(models11_1)

Call:
lm(formula = score1 ~ special + free + stratio + income + noteng,
    data = train_edata2)

Residuals:
    Min      1Q  Median      3Q     Max
-20.642  -4.660  -1.088   4.766  23.919

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 721.14713    7.67740  93.931  < 2e-16 ***
special      -0.41759    0.20244  -2.063   0.0408 *
free         -0.52269    0.06632  -7.882 5.79e-13 ***
stratio      -0.67239    0.31646  -2.125   0.0352 *
income        0.82671    0.15783   5.238 5.33e-07 ***
noteng       -0.47898    0.27559  -1.738   0.0842 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.291 on 152 degrees of freedom
Multiple R-squared:  0.7378,    Adjusted R-squared:  0.7291
F-statistic: 85.52 on 5 and 152 DF,  p-value: < 2.2e-16
```

Residuals plot didn’t show any strange pattern and the model residuals are spare and randomly distributed around 0 (Figure 2.8). C+R Plots highlighted the necessity of a quadratic transformation over the **income** predictor (Figure 2.9).

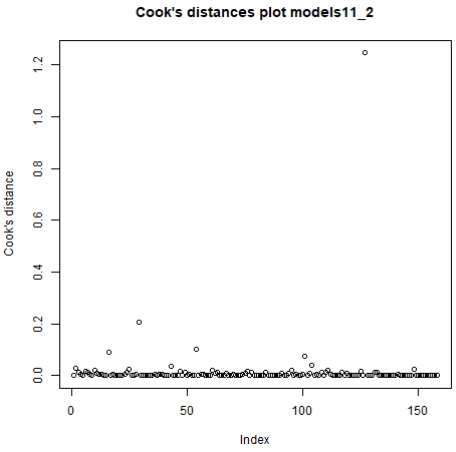
Figure 2.8 Figure 2.9



A new **model11_2** was fit considering the quadratic transformation (Figure 2.10). All the variables were significant - the model explained 73.6% of the variability of the response **score1**. Plotting the cook distances of the model we can spot the presence of a high leverage point (Figure 2.11). After removal of the high leverage point the clean model **models11_3** was fit. With the removal of the high leverage point, significance of the quadratic transformation was lost for the variable **income**. Model **model11_4** was fit without the quadratic transformation.

Figure 2.10 Figure 2.11

```
> summary(models11_2)

Call:
lm(formula = score1 ~ special + free + stratio + income + noteng +
    I(income^2), data = train_edata2)

Residuals:
    Min      1Q  Median      3Q     Max
-21.016  -4.593  -1.014   3.983  24.310

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 704.17731   10.91419  64.519  < 2e-16 ***
special      -0.44943    0.20057  -2.241  0.02650 *
free         -0.43810    0.07631  -5.741 5.01e-08 ***
stratio      -0.71208    0.31323  -2.273  0.02442 *
income        2.35621    0.72421   3.253  0.00141 **
noteng       -0.57843    0.27617  -2.094  0.03789 *
I(income^2)  -0.03052    0.01411  -2.163  0.03214 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.193 on 151 degrees of freedom
Multiple R-squared:  0.7456,    Adjusted R-squared:  0.7355
F-statistic: 73.77 on 6 and 151 DF,  p-value: < 2.2e-16
```

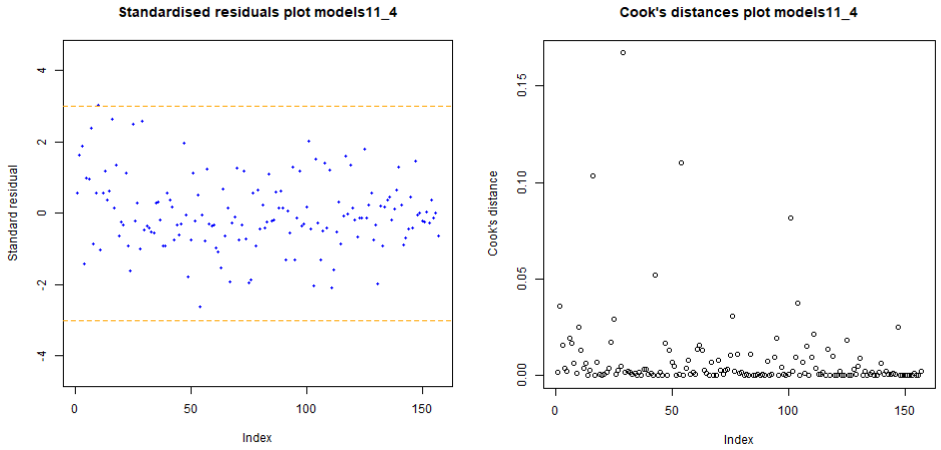


Model **model11_4** was chosen as our best model, it explains 74.0% of the variability of the response **score1** (Figure 2.12). Residual standard error is 8.14, a value that can be considered acceptable if related to the average of the response **score 1** (708.8). Standard residual plot shows residuals randomly sparce around zero and no pattern is present, also, there are no outliers (Figure 2.13). Cook's distances plot excludes the presence of high leverage points (Figure 2.14).

Figure 2.12

```
> summary(models11_4)

Call:
lm(formula = score1 ~ special + free + stratio + noteng + income,
    data = train_edata2_c)

Residuals:
    Min      1Q  Median      3Q     Max
-20.319  -4.565  -1.042   4.487  24.345

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 716.59571    7.73064  92.695  < 2e-16 ***
special      -0.42321    0.19865  -2.130   0.0348 *
free         -0.47795    0.06727  -7.105 4.44e-11 ***
stratio      -0.67513    0.31052  -2.174   0.0312 *
noteng       -0.55420    0.27193  -2.038   0.0433 *
income        1.05740    0.17812   5.937 1.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.136 on 151 degrees of freedom
Multiple R-squared:  0.7478,    Adjusted R-squared:  0.7395
F-statistic: 89.56 on 5 and 151 DF,  p-value: < 2.2e-16
```

Figure 2.13 Figure 2.14



Model fit using **score2** as the response.

A full model **MLRs16** was fit to assess preliminary model performance when all the variables are included (Figure 2.15).

Figure 2.15

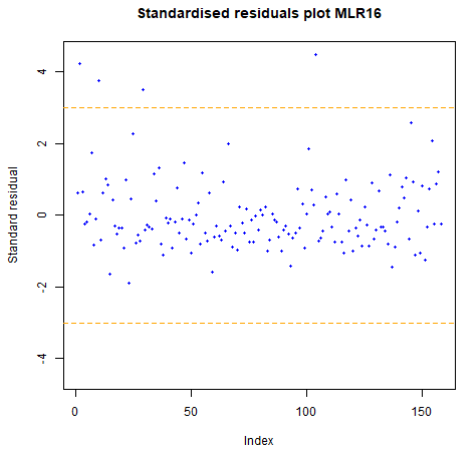
```
Call:
lm(formula = score2 ~ . - ID - score1, data = train_edata2)

Residuals:
    Min      1Q  Median      3Q     Max
-20.540  -7.753  -3.119   5.396  51.702

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.090e+02  1.725e+01  41.101  < 2e-16 ***
spendspec    1.635e-03  8.129e-04   2.011   0.0462 *
spendpp     -2.835e-03  2.083e-03  -1.362   0.1754
ipad        -5.836e-01  3.849e-01  -1.516   0.1316
special     -1.637e-01  3.297e-01  -0.496   0.6204
free        -6.694e-01  1.059e-01  -6.322 2.88e-09 ***
stratio     -1.176e+00  5.668e-01  -2.074   0.0398 *
income       1.536e+00  3.135e-01   4.900 2.49e-06 ***
salary       7.582e-02  4.207e-01   0.180   0.8572
noteng      -2.817e-01  4.309e-01  -0.654   0.5143
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.43 on 148 degrees of freedom
Multiple R-squared:  0.7111,    Adjusted R-squared:  0.6935
F-statistic: 40.47 on 9 and 148 DF,  p-value: < 2.2e-16
```

From the full model we can see that only the **intercept** and the variables **free**, **spendspec**, **income** and **stratio** are significant. Presence of collinearity was excluded with by Variance Inflator Factor analysis (VIF)(Figure 2.16). Standard residuals plot shows the presence of outliers in the model (Figure 2.17).

Figure 2.16

```
> vif(MLRs16)
spendspec   spendpp      ipad   special      free   stratio    income    salary    noteng
 2.231102  4.183718  1.130487  1.277613  3.207244  1.609785  2.781961  2.004382  2.100220
```

Figure 2.17



The outliers are removed and full model **MLRs16cc** is fit. Step AIC analysis (both forward and backward) was performed on the model and it let to the following best fit models:

Forward stepAIC: score2 ~ free + income + spendspec, AIC = 683.29

Backward stepAIC: score2 ~ special + free + stratio + income, AIC = 667.78

Model from backwards stepAIC was selected as the best model due to a considerably lower AIC score. Model **model16_p1** was fit (Figure 2.18). The fitted model presents an adjusted-R^2 of 0.8301, however two of the predictors are not statistically significative (**special** and **stration**) and will be removed from the model.

Figure 2.18

```
> summary(model16_p1)

Call:
lm(formula = score2 ~ income + free + special + stratio, data = train_edata2_ff)

Residuals:
    Min      1Q  Median      3Q     Max
-18.676  -5.890  -1.905   5.074  27.222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 692.36962    8.53624  81.109   <2e-16 ***
income        1.76753    0.16664  10.607   <2e-16 ***
free         -0.70856    0.05438 -13.029   <2e-16 ***
special      -0.33543    0.22241  -1.508    0.134
stratio      -0.57616    0.35107  -1.641    0.103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.851 on 147 degrees of freedom
Multiple R-squared:  0.8346,    Adjusted R-squared:  0.8301
F-statistic: 185.5 on 4 and 147 DF,  p-value: < 2.2e-16
```

Model **model16_p2** was fit (Figure 2.19). It explains 82.8% of the variability of the response, standard residuals plot shows no outliers, and the residues are sparsely distributed around 0 and they show no clear pattern (Figure 2.20). There is no presence of high leverage points and the C+R plots are ideal (Figure 2.21, Figure 2.22).

Figure 2.19

```
> summary(model16_p2)

Call:
lm(formula = score2 ~ income + free, data = train_edata2_ff)

Residuals:
    Min      1Q  Median      3Q     Max
-20.457  -5.845  -1.363   5.118  24.749

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 676.92957    3.71054  182.43   <2e-16 ***
income        1.79927    0.16688   10.78   <2e-16 ***
free         -0.73823    0.05275  -13.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.914 on 149 degrees of freedom
Multiple R-squared:   0.83,     Adjusted R-squared:  0.8277
F-statistic: 363.7 on 2 and 149 DF,  p-value: < 2.2e-16
```

Figure 2.20

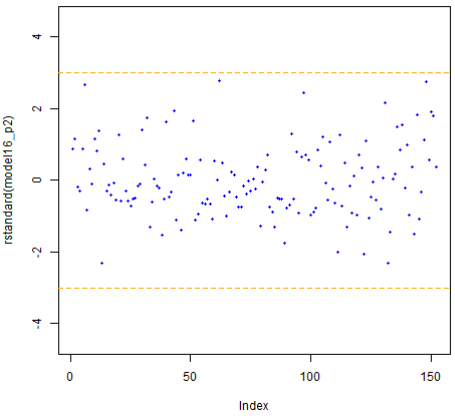


Figure 2.21

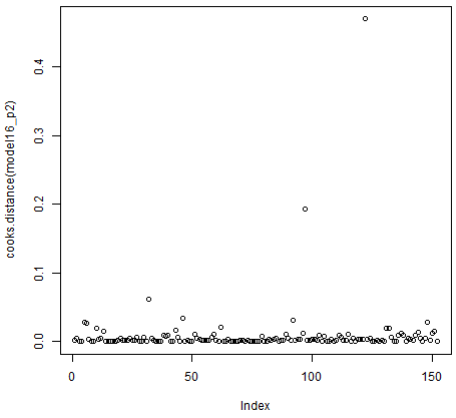
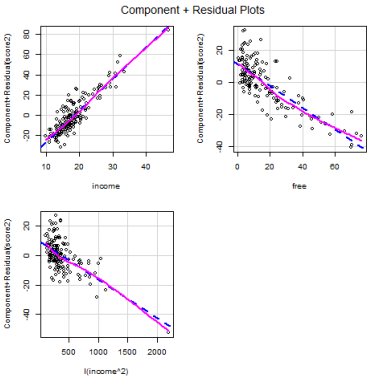


Figure 2.22



Best model:

score1 = 716.596 - 0.423 **special** - 0.478 **free** - 0.675 **stratio** - 0.55 **noteng** + 1.057 **income**

score2 = 676.930 + 1.799 **income** - 0.738 **free**

Model interpretation:

The linear regression fit for the two models leads to the conclusion that income has by far the biggest influence over both the responses **score1** and **score 2**, in case of **score 1** an increase of 1 (x1000€) of income per student by the school increases by 1.057 the average score, while it is increases by almost double of it (1.799) considering **score 2**. Considering the best model fit for **score 1**, there is a negative influence on the response in case high percentages of students with special needs (**special**) and in case a big percentage of students eligible for free lunch (**free**), this last value might be linked to an increased percentage of disadvantaged students. Also, a high student to teacher ratio (**stratio**) has a negative influence on the response, lowering the average response by 0.675, also, an high percentage of students which first language is not English in the classes has a negative effect on the response. However, considering **score2**, only **income** and **free** have a considerable influence on the response.

Model evaluation:

model11_4

model16_p2

Predictions were made using the data of the test dataset formed of 100 observations – model **model11_4** and **model16_p2** were used to predict the responses of **score1** and **score2**. Performance of the models was evaluated with as the mean absolute percentage error (MAPE).

$$MAPE = 100 - \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Actual_t - Forecast_t}{Actual_t}\right|$$

A MAPE value between 0% and 10%, accounts for very high forecasting ability of the model.

model11_4 obtained a MAPE of 1.1%, highlighting high forecast ability of **score1**.

model16_p2 obtained a MAPE of 1.6%, highlighting high forecast ability of **score2**.

References

[1] NIH – National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9959609 (accessed on April 2024)