LIS542 Midterm - Replicating a study

Jialu Wang(jwang282)

Background:

This paper refers to “A quantitative model for linking two disparate sets of articles in MEDLINE” (Torvik and Smalheiser, 2007), replicate the S2 model in the article, analyze the data, propose a new logistic regression, test the fitness of the new mode and

Step 1: Read the paper and learned about literature-based discovery.

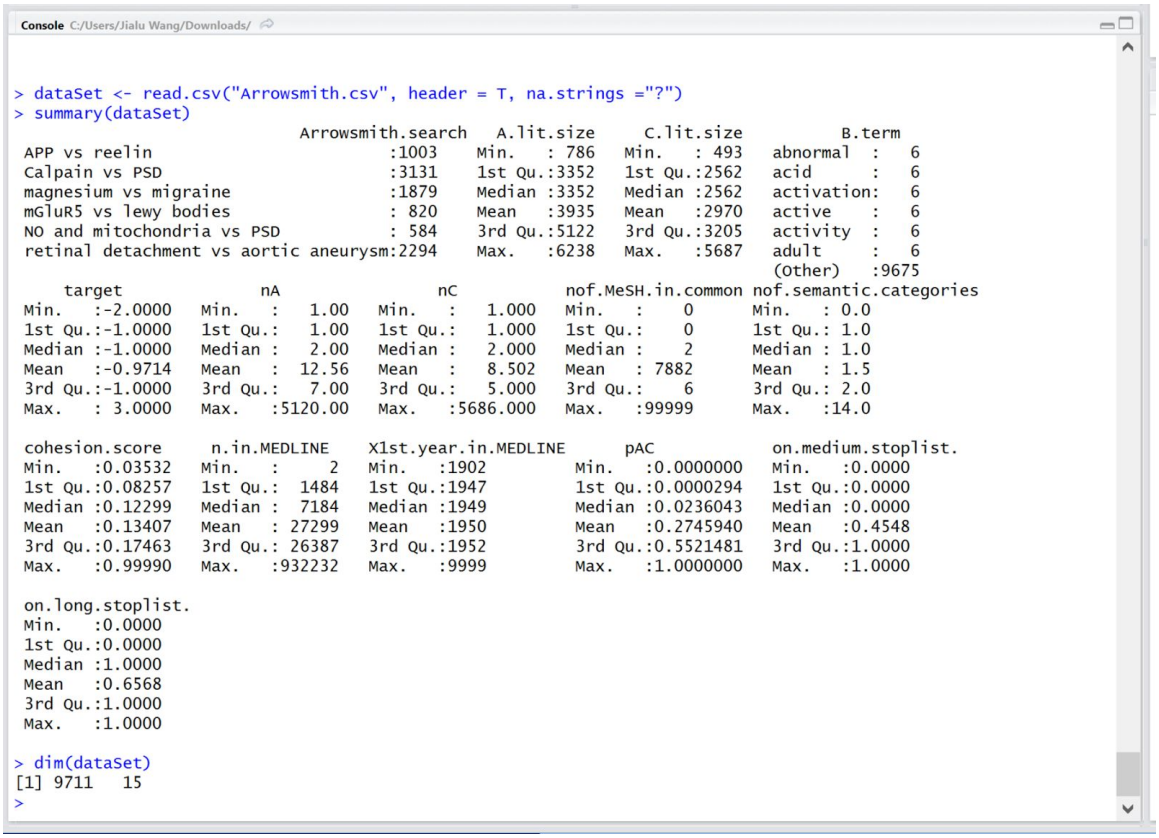
Step 2: Convert the Excel file (.xls) to comma-separated-value plain text file (.csv).

Save the file “Arrowsmith.xls” as csv to convert it. For convenience of further procedures, I delete the first four sentences of the file.

Step 3: Load the csv file into R and construct some attributes.

R code of this step is attached in the R file.

The screenshot of loading dataset into Rstudio is as follow:



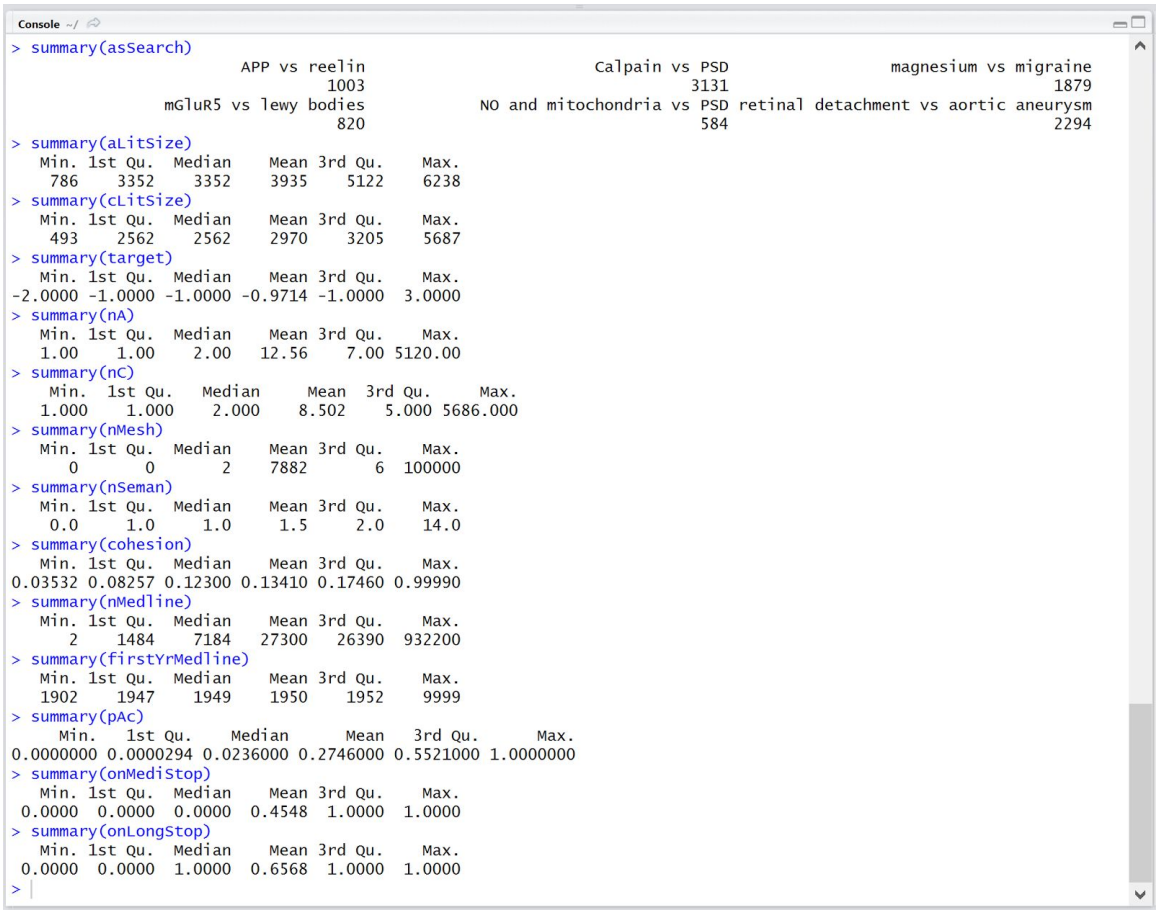
```
Console C:/Users/Jialu Wang/Downloads/

> dataSet <- read.csv("Arrowsmith.csv", header = T, na.strings ="?")
> summary(dataSet)
                          Arrowsmith.search   A.lit.size     C.lit.size         B.term
 APP vs reelin                       :1003    Min.   : 786   Min.   : 493   abnormal  :   6
 Calpain vs PSD                      :3131    1st Qu.:3352   1st Qu.:2562   acid      :   6
 magnesium vs migraine               :1879    Median :3352   Median :2562   activation:   6
 mGluR5 vs lewy bodies               : 820    Mean   :3935   Mean   :2970   active    :   6
 NO and mitochondria vs PSD          : 584    3rd Qu.:5122   3rd Qu.:3205   activity  :   6
 retinal detachment vs aortic aneurysm:2294    Max.   :6238   Max.   :5687   adult     :   6
                                                                           (Other)   :9675
    target             nA                nC           nof.MeSH.in.common nof.semantic.categories
 Min.   :-2.0000   Min.   :   1.00   Min.   :   1.000   Min.   :    0      Min.   : 0.0
 1st Qu.:-1.0000   1st Qu.:   1.00   1st Qu.:   1.000   1st Qu.:    0      1st Qu.: 1.0
 Median :-1.0000   Median :   2.00   Median :   2.000   Median :    2      Median : 1.0
 Mean   :-0.9714   Mean   :  12.56   Mean   :   8.502   Mean   : 7882      Mean   : 1.5
 3rd Qu.:-1.0000   3rd Qu.:   7.00   3rd Qu.:   5.000   3rd Qu.:    6      3rd Qu.: 2.0
 Max.   : 3.0000   Max.   :5120.00   Max.   :5686.000   Max.   :99999      Max.   :14.0

 cohesion.score     n.in.MEDLINE    X1st.year.in.MEDLINE      pAC            on.medium.stoplist.
 Min.   :0.03532   Min.   :     2   Min.   :1902         Min.   :0.0000000   Min.   :0.0000
 1st Qu.:0.08257   1st Qu.:  1484   1st Qu.:1947         1st Qu.:0.0000294   1st Qu.:0.0000
 Median :0.12299   Median :  7184   Median :1949         Median :0.0236043   Median :0.0000
 Mean   :0.13407   Mean   : 27299   Mean   :1950         Mean   :0.2745940   Mean   :0.4548
 3rd Qu.:0.17463   3rd Qu.: 26387   3rd Qu.:1952         3rd Qu.:0.5521481   3rd Qu.:1.0000
 Max.   :0.99990   Max.   :932232   Max.   :9999         Max.   :1.0000000   Max.   :1.0000

 on.long.stoplist.
 Min.   :0.0000
 1st Qu.:0.0000
 Median :1.0000
 Mean   :0.6568
 3rd Qu.:1.0000
 Max.   :1.0000

> dim(dataSet)
[1] 9711   15
>
```

Variables transformed are as follow:

Variables	Name
asSearch	Arrowsmith search
aLitSize	A-lit size
cLitSize	C-lit size
target	target
nA	nA
nC	nC
nMesh	nof MeSH in common
nSeman	nof semantic categories
cohesion	cohesion score
nMedline	n in MEDLINE
firstYrMedline	1st year in MEDLINE
pAc	pAC
onMediStop	on medium stoplist?
onLongStop	on long stoplist?

Step 4: Get to know the dataset: assess the summary statistics, histograms, and pairwise scatter plots before and after your transformation. Are there missing values or outliers?

Statistics



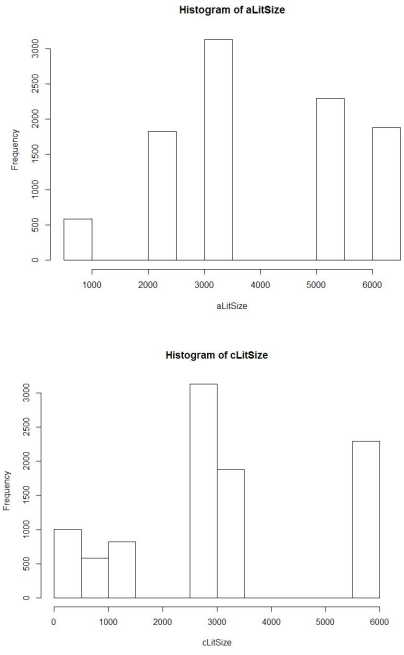
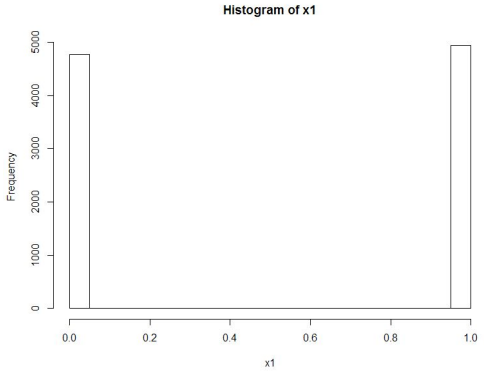
```
Console ~/
> summary(asSearch)
                        APP vs reelin                         Calpain vs PSD                 magnesium vs migraine
                                 1003                                  3131                                  1879
                mGluR5 vs lewy bodies            NO and mitochondria vs PSD retinal detachment vs aortic aneurysm
                                  820                                   584                                  2294
> summary(aLitSize)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    786    3352    3352    3935    5122    6238
> summary(cLitSize)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    493    2562    2562    2970    3205    5687
> summary(target)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.0000 -1.0000 -1.0000 -0.9714 -1.0000  3.0000
> summary(nA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    2.00   12.56    7.00 5120.00
> summary(nC)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   1.000    1.000    2.000    8.502    5.000 5686.000
> summary(nMesh)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0       2    7882       6  100000
> summary(nSeman)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     1.0     1.0     1.5     2.0    14.0
> summary(cohesion)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03532 0.08257 0.12300 0.13410 0.17460 0.99990
> summary(nMedline)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2    1484    7184   27300   26390  932200
> summary(firstYrMedline)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1902    1947    1949    1950    1952    9999
> summary(pAc)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000000 0.0000294 0.0236000 0.2746000 0.5521000 1.0000000
> summary(onMediStop)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4548  1.0000  1.0000
> summary(onLongStop)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.6568  1.0000  1.0000
>
```

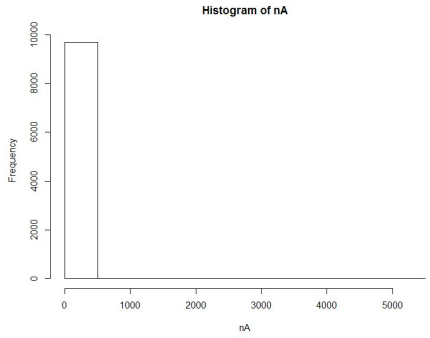
Before

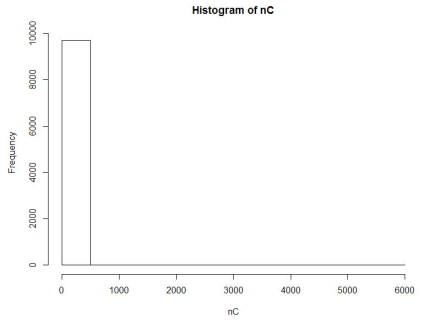
```
Console ~/
> summary(x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5092  1.0000  1.0000
> summary(x2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   0.661   1.000   1.000
> summary(x3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   0.788   1.000   1.000
> summary(x4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03532 0.08257 0.12300 0.13350 0.17460 0.30000
> summary(x5)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.9700000 -1.4630000 -0.9739000 -1.0120000 -0.4933000 -0.0004341
> summary(x6)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1950    1950    1950    1955    1952    2005
> summary(x7)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2579  1.6270  2.7400  4.5320  8.0000
> summary(I1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2362  0.0000  1.0000
> summary(I2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.06014 0.00000 1.00000
> summary(I3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.08444 0.00000 1.00000
> summary(I4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1935  0.0000  1.0000
> summary(I5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3224  1.0000  1.0000
> summary(I6)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1033  0.0000  1.0000
> summary(I6)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1033  0.0000  1.0000
```

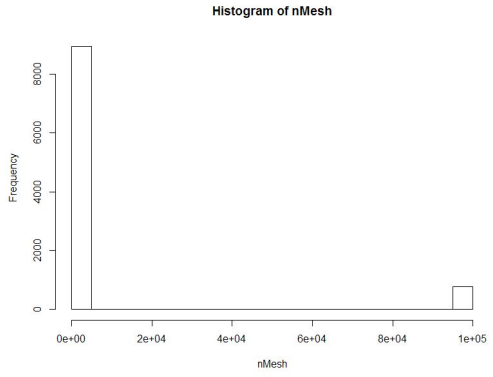
After

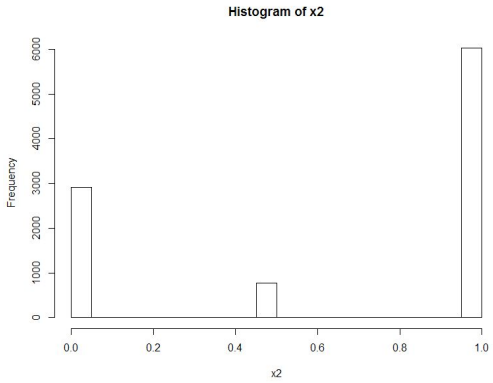
Histograms

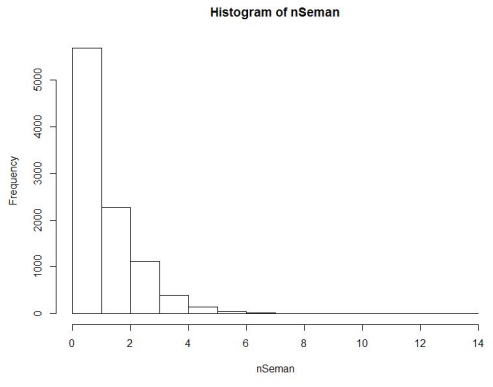
Before	After
	

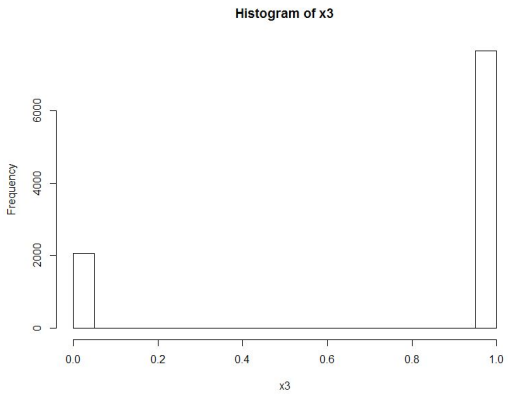

Histogram of nA
Frequency
0
2000
4000
6000
8000
10000
0
1000
2000
3000
4000
5000
nA

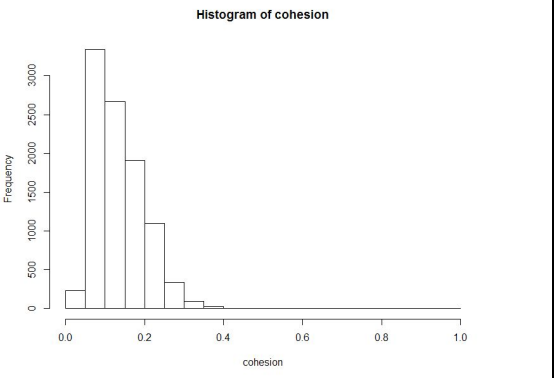

Histogram of nC
Frequency
0
2000
4000
6000
8000
10000
0
1000
2000
3000
4000
5000
6000
nC

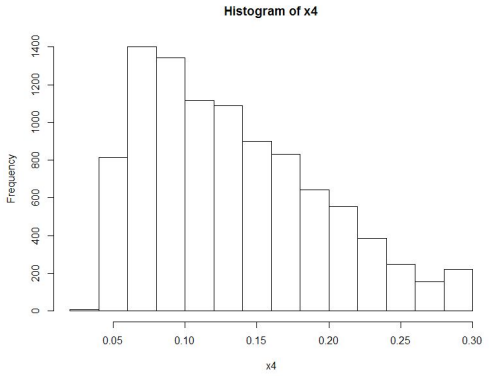

Histogram of nMesh
Frequency
0
2000
4000
6000
8000
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
nMesh

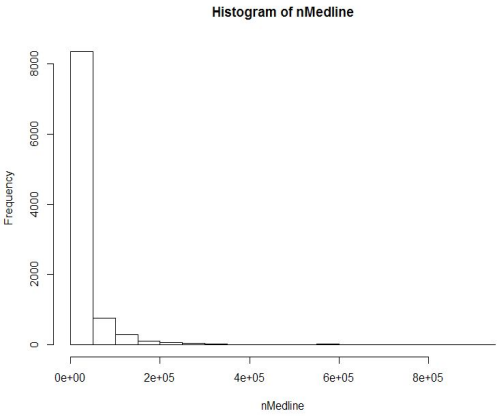

Histogram of x2
Frequency
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
1.0
x2

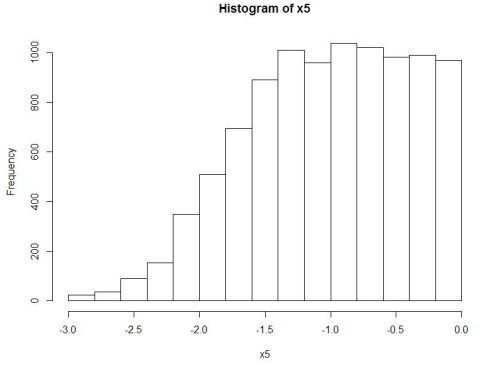

Histogram of nSeman
Frequency
0
1000
2000
3000
4000
5000
0
2
4
6
8
10
12
14
nSeman

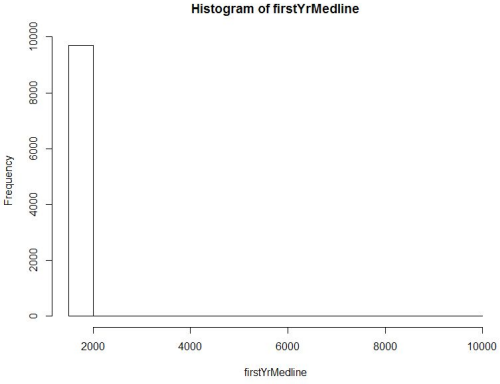

Histogram of x3
Frequency
0
2000
4000
6000
0.0
0.2
0.4
0.6
0.8
1.0
x3

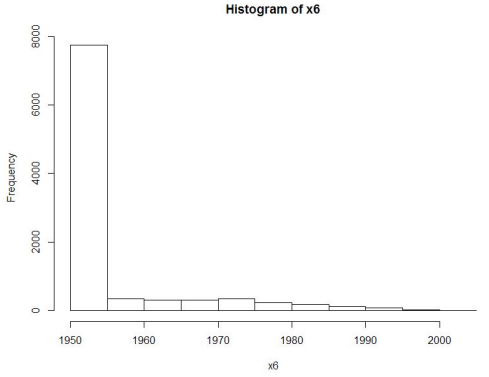

Histogram of cohesion
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
cohesion

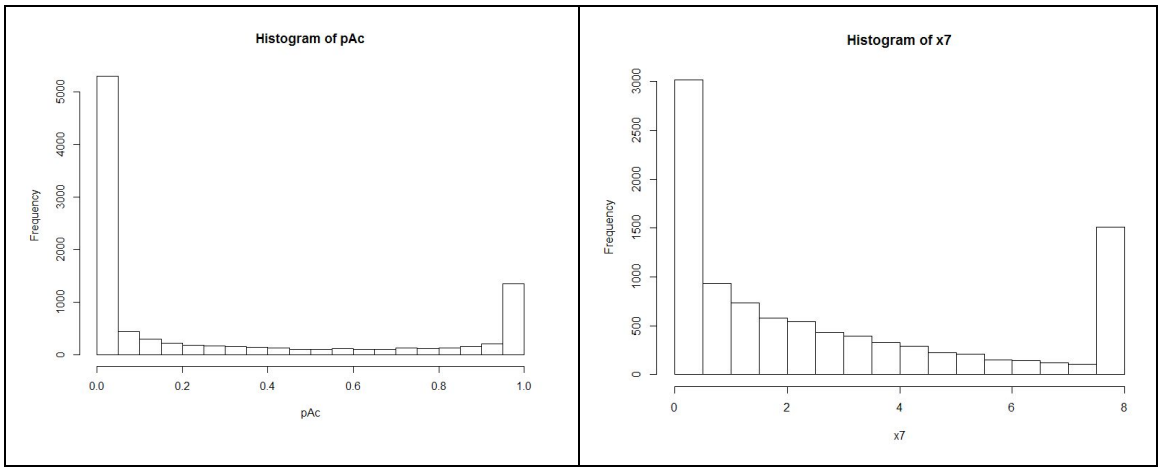

Histogram of x4
Frequency
0
200
400
600
800
1000
1200
1400
0.05
0.10
0.15
0.20
0.25
0.30
x4


Histogram of nMedline
Frequency
0
2000
4000
6000
8000
0e+00
2e+05
4e+05
6e+05
8e+05
nMedline

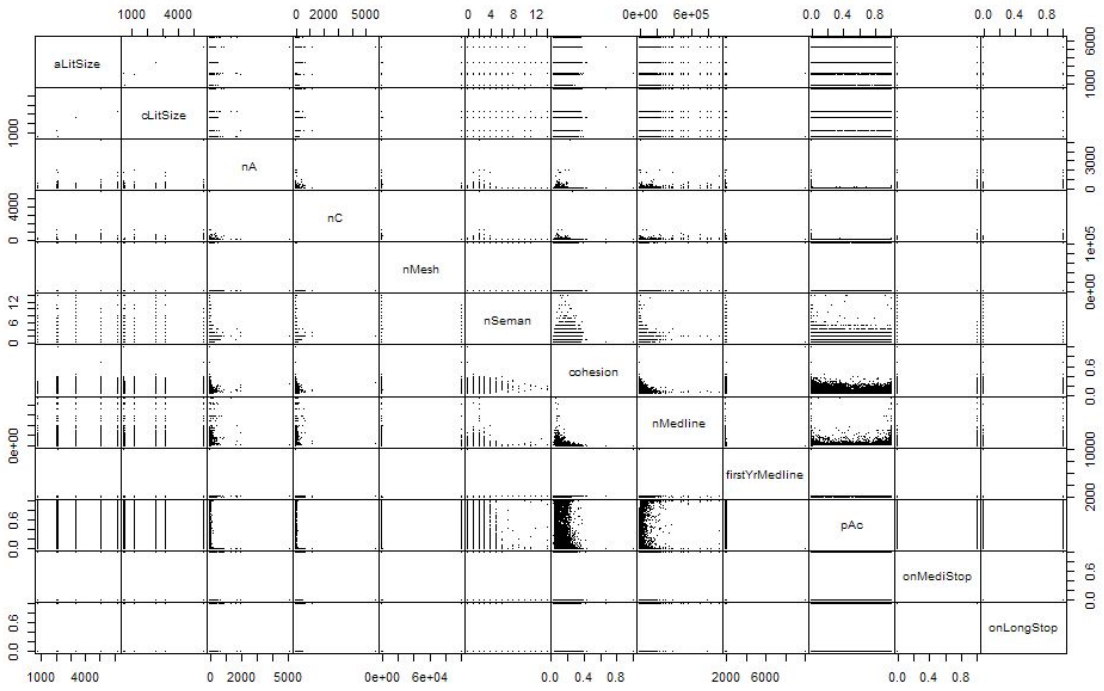

Histogram of x5
Frequency
0
200
400
600
800
1000
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
x5


Histogram of firstYrMedline
Frequency
0
2000
4000
6000
8000
10000
2000
4000
6000
8000
10000
firstYrMedline

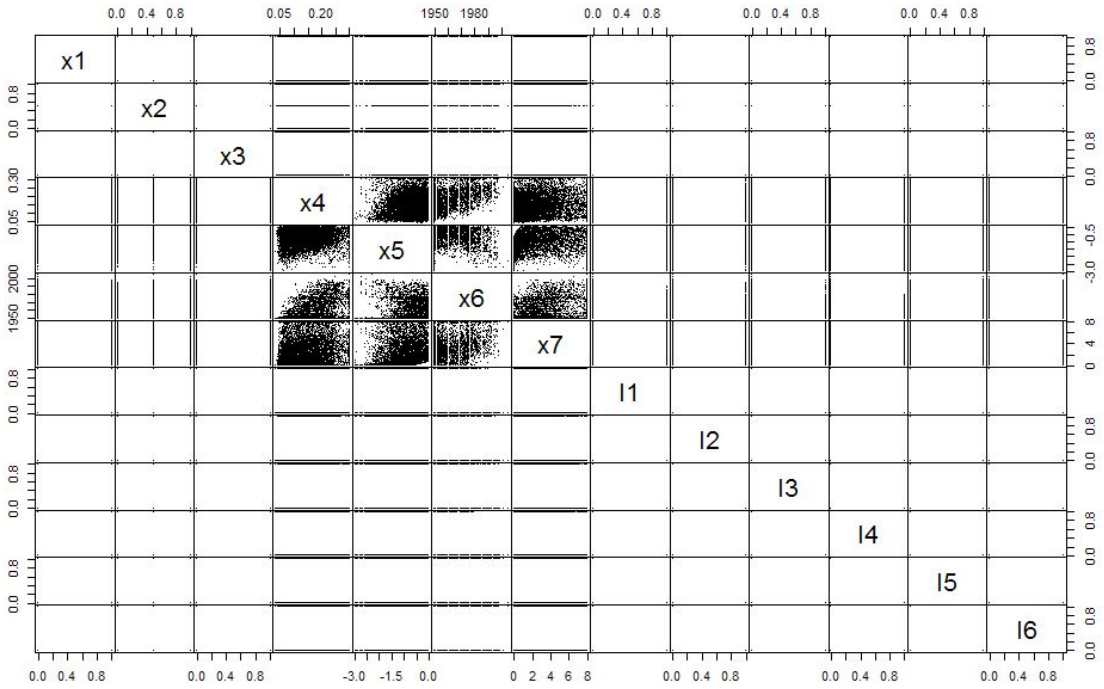

Histogram of x6
Frequency
0
2000
4000
6000
8000
1950
1960
1970
1980
1990
2000
x6



Pairwise scatter plots



Before



After

Findings:

1) nA outlier:

Arrowsmith search	A-lit size	C-lit size	B-term	target	nA	nC
retinal detachment vs aortic aneurysm	5122	5687	detachment	-1	5120	1

2) nC outlier:

Arrowsmith search	A-lit size	C-lit size	B-term	target	nA	nC
retinal detachment vs aortic aneurysm	5122	5687	aneurysm	-1	1	5686

3) *nof MeSH in common* missing value:

those with value 99999, 765 in total

4) *cohesion* and *1st year in MEDLINE* missing value:

Arrowsmith search	A-lit size	C-lit size	B-term	n A	n C	nof MeSH in common	cohesion score	1st year in MEDLINE
APP vs reelin	2118	493	receptor apoer2	1	2	99999	0.9999	9999

5) Scatterplots before transformation:

not much inferences can be made.

6) Scatterplots after transformation:

it seems that x4 and x7, x5 and x7, x6 and x7 are independent, x4 and x5, x4 and x6, x5 and x6 have some kind of relationship.

Step 5: Fit a logistic regression model and assess the validity of its assumptions and statistical significance. Interpret the parameters and your model. Are your parameter estimates different from the ones reported? If so, why?

As the p-values are really small – assumptions statistically at significant level 99.9%, except for i4 which is statistically significant at level 99%.

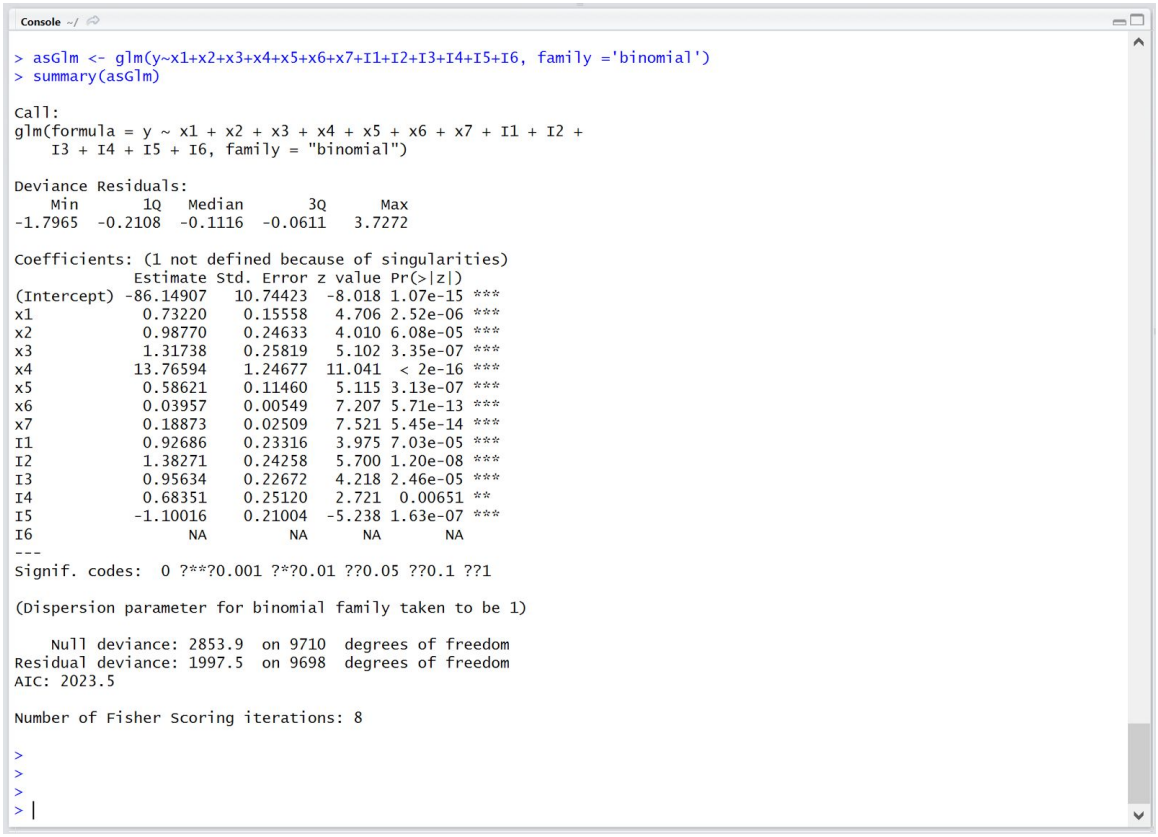
Interpretation of parameters: Each x_i increase 1, y will increase by the corresponding parameter.

My model: y = 0.732x1 + 0.988x2 + 1.317x3 + 13.766x4 + 0.586x5 + 0.0396x6 + 0.189x7

S2 model:

$$\text{B-term score} : y = 0.73x_1 + 0.99x_2 + 1.32x_3 + 13.8x_4 + 0.59x_5 + 0.040x_6 + 0.19x_7.$$

So my estimates for parameters x1 – x7 are almost the same with those in the S2 model.



```
Console ~/

> asGlm <- glm(y~x1+x2+x3+x4+x5+x6+x7+I1+I2+I3+I4+I5+I6, family ='binomial')
> summary(asGlm)

Call:
glm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + I1 + I2 +
    I3 + I4 + I5 + I6, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7965  -0.2108  -0.1116  -0.0611   3.7272

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -86.14907   10.74423  -8.018 1.07e-15 ***
x1            0.73220    0.15558   4.706 2.52e-06 ***
x2            0.98770    0.24633   4.010 6.08e-05 ***
x3            1.31738    0.25819   5.102 3.35e-07 ***
x4           13.76594    1.24677  11.041  < 2e-16 ***
x5            0.58621    0.11460   5.115 3.13e-07 ***
x6            0.03957    0.00549   7.207 5.71e-13 ***
x7            0.18873    0.02509   7.521 5.45e-14 ***
I1            0.92686    0.23316   3.975 7.03e-05 ***
I2            1.38271    0.24258   5.700 1.20e-08 ***
I3            0.95634    0.22672   4.218 2.46e-05 ***
I4            0.68351    0.25120   2.721  0.00651 **
I5           -1.10016    0.21004  -5.238 1.63e-07 ***
I6                 NA         NA      NA       NA
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2853.9  on 9710  degrees of freedom
Residual deviance: 1997.5  on 9698  degrees of freedom
AIC: 2023.5

Number of Fisher Scoring iterations: 8

>
>
>
>
```

Step 6. Reflect on aspects that made the process easy or hard.

1) Why I6 has null value?

A possible reason I6 having null value may be related to the term "overfitting". Overfitting occurs in excessively complex statistical models, with too many parameters relative to the number of observations. It usually indicates that the criterion used for training the model is not the same as the criterion used to judge the efficacy of a model. In this model, glm is not able to summaries all the indicators perfectly because of overfitting. Therefore, the last indicator becomes NA.

To avoid overfitting, additional techniques such as cross-validation should be applied to penalize the over-complex model. This could be implemented in future work.

2) Testing the dependence of x4, x5, x6.

pair wise scatter plot can be performed to discover the relationships among x4, x5 and x6. From a scratch of the plots we could roughly recognize a proportional relationship between x4 and x6, as well as an inversely proportional between x5 and x6. More statistical analysis could be made in future work.

Conclusion:

In this report, a replica of S2 model based on the topic "A quantitative model for linking two disparate sets of articles in MEDLINE" is built by constructing a generalized linear model. The result derived from the new model is y = 0.732x1 + 0.988x2 + 1.317x3 + 13.766x4 + 0.586x5 + 0.0396x6 + 0.189x7, which is close to the result in the original model. Several dependencies between the variables are found from the regression model as well for further illustration.

The hardest part of the project is analyzing the problem. It took time to understand the background and objective of the project since the original model was built from a study in an unfamiliar field. After analyzed the problem clearly, it is relatively simple to follow the steps from the project instruction.

References:

Anon. Douglas M. Hawkins, School of Statistics, University of Minnesota - Research. Retrieved October 31, 2016 from http://users.stat.umn.edu/~dhawkins/research/research.htm

Vetle I.Torvik and. Vetle I. Torvik. Retrieved October 31, 2016 from http://bioinformatics.oxfordjournals.org/content/23/13/1658.full