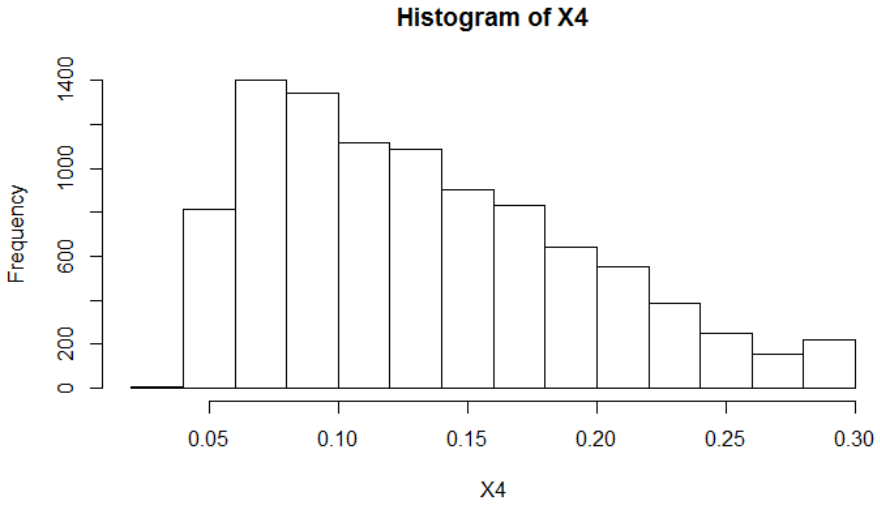
```
# LIS542 Midterm
# Hui Lyu

# Replicating a study

###################################
### Step 1: Load csv file into R

arrow = read.csv("C:/Documents/GSLIS/542 Data, Statistics and Information/Midterm/Arrowsmith.csv", header = TRUE, skip = 4)
arrow[,c(16:29)] = NULL
dim(arrow)
# [1] 9711   15
# There are 9711 records in total.

###################################
### Step 2: Construct attributes
X1 = ifelse(((arrow$nA>1 | arrow$A.lit.size<1000) & (arrow$nC>1 | arrow$C.lit.size<1000)) , 1, 0)

X2 = ifelse((arrow$nof.MeSH.in.common>0 & arrow$nof.MeSH.in.common<99999), 1, ifelse(arrow$nof.MeSH.in.common==99999, 0.5, 0))

X3 = ifelse(arrow$nof.semantic.categories>0, 1, 0)

X4 = ifelse(arrow$cohesion.score<0.3, arrow$cohesion.score, 0.3)

X5 = -abs(log10(arrow$n.in.MEDLINE)-3)

X6.1 = ifelse(arrow$X1st.year.in.MEDLINE<2005, arrow$X1st.year.in.MEDLINE, 2005)
X6 = ifelse(X6.1>1950, X6.1, 1950)

X7 = ifelse(8 < -log10(arrow$pAC+0.000000001), 8, -log10(arrow$pAC+0.000000001))
```

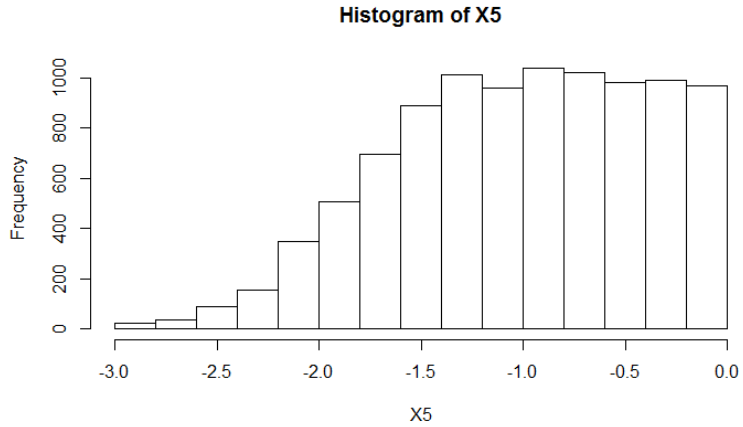
```
I1 = ifelse(arrow$Arrowsmith.search=="retinal detachment vs aortic aneurysm", 1, 0)

I2 = ifelse(arrow$Arrowsmith.search=="NO and mitochondria vs PSD", 1, 0)

I3 = ifelse(arrow$Arrowsmith.search=="mGluR5 vs lewy bodies", 1, 0)

I4 = ifelse(arrow$Arrowsmith.search=="magnesium vs migraine", 1, 0)

I5 = ifelse(arrow$Arrowsmith.search=="Calpain vs PSD", 1, 0)

I6 = ifelse(arrow$Arrowsmith.search=="APP vs reelin", 1, 0)

Y = ifelse((arrow$target==0 | arrow$target==2), 1, 0)


##################################
### Step 3: Assess the dataset

### summary statistics
sum(X1==1)
# [1] 4945
sum(X1==0)
# [1] 4766

# X1 is set to select B-terms that occur in more than one paper with literatures A and C, and
# B-terms that satisfy the expression referring to the literature size. The values of X1 of those
# B-terms are 1.

sum(X2==1)
# [1] 6036
sum(X2==0)
# [1] 2910
```

```
sum(X2==0.5)
# [1] 765

# X2 is set to distinguish B-terms based on number of MeSH in common. Since we set the value of
# 99999 to the missing values, there is an interval partition at value 99999. We are not sure
# whether B-terms with missing values in number of MeSH in common are relevant, so X2 is set to 0.5
# (the median value between 0 and 1) for those B-terms.

sum(X3==1)
# [1] 7652
sum(X3==0)
# [1] 2059

# X3 is set to select B-terms that map to at least one UMLS semantic category, and set them value
# 1. Others are set to value 0. We can find that most B-terms map one or more semantic categories.

summary(X4)
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.03532 0.08257 0.12300 0.13350 0.17460 0.30000

# X4 is set to distinguish B-terms based on cohesion score. Scores greater than 0.3 is set to 0.3
# It can avoid nonlinearities observed at very high values.

summary(X5)
#       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
# -2.9700000 -1.4630000 -0.9739000 -1.0120000 -0.4933000 -0.0004341

# X5 is set to select B-terms that are not extremely common or extremely rare in general usage.
# Very frequent or very infrequent terms are penalized correspondingly.


summary(X6)
```

```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1950    1950    1950    1955    1952    2005

# X6 is set to restrict B-terms appearing year between 1950 and 2005.

summary(X7)
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.0000  0.2579  1.6270  2.7400  4.5320  8.0000

# X7 implies that the more characteristic the B-term is within A or C, the higher its score.

### histograms
# There is no need to plot histograms of X1, X2 and X3. The distributions can be obtained in the
# summary function of the former step.

hist(X4)
```

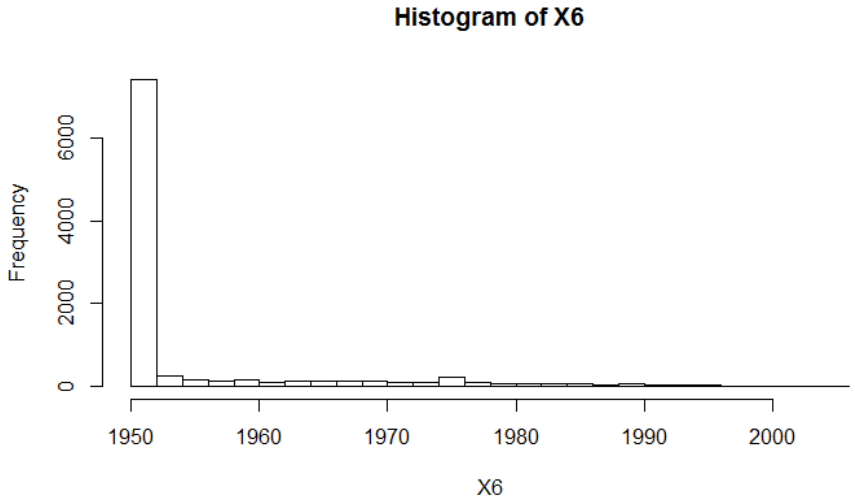



```
# The distribution of X4 is nearly right skewed.

hist(X5)
```



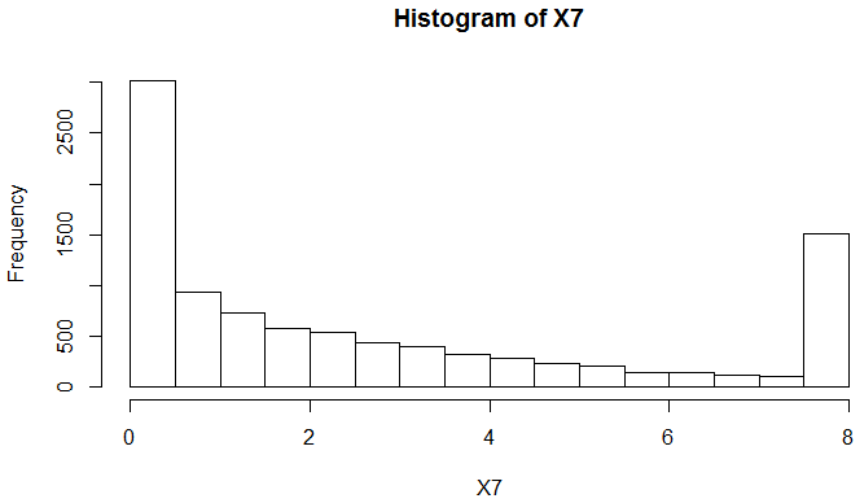
The distribution of X5 is nearly left skewed. While most B-terms have values between -1.5 and 0.

hist(X6, breaks = 30)



Most B-terms have values of 1950.

hist(X7)



The distribution is nearly right skewed. Since we set number 8 as a cut off, so there are more

B-terms have value 8 for X7.

pairwise scatter plots before and after transformation

For X1, X2 and X3, the concrete numbers have very large ranges also including missing values

and outliers. So these transformations help to select and classify the B-terms, and set the

values to 1 or 0 for good statistical filter and further use of logistic regression model.

table(arrow$nof.MeSH.in.common)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2910	1922	1034	649	467	297	232	171	142	120	112	82	61	45	50	55
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
30	47	34	21	20	26	17	18	21	14	9	14	8	15	18	11
32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
12	8	12	12	7	5	7	8	7	6	4	7	2	3	6	6
48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
1	6	4	9	3	2	2	3	7	2	2	4	2	2	2	8
64	66	67	68	69	70	71	73	74	75	77	78	80	81	85	86
3	2	6	1	4	3	4	2	1	1	1	1	4	1	2	1
87	88	89	91	93	94	95	97	98	99	100	103	104	105	109	110
1	2	1	2	1	1	1	1	2	1	1	1	1	2	1	1
113	115	116	120	122	123	124	125	129	131	133	136	138	139	145	148
1	1	1	1	2	1	1	1	2	1	1	1	3	2	1	2
150	152	154	156	158	163	164	172	177	178	186	194	195	198	200	211
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
223	232	240	241	299	321	388	509	99999							
1	1	1	1	1	1	1	1	765							

For example, the original values of number of MeSH in common are messy. The transformation makes

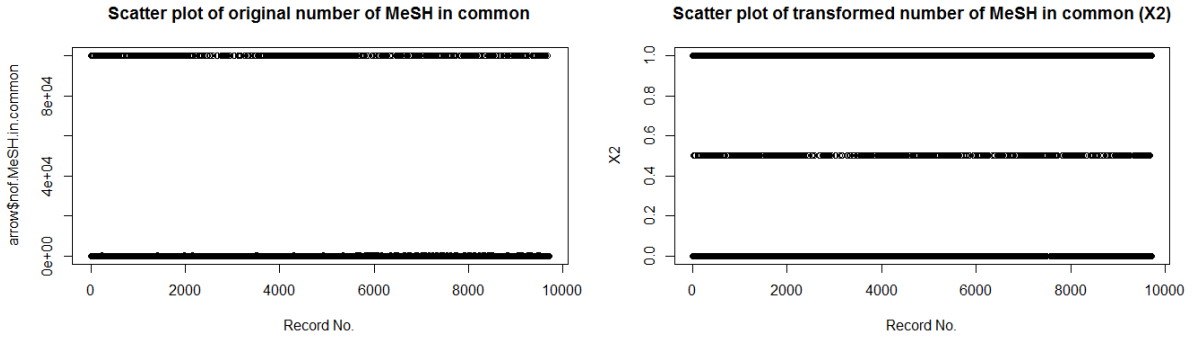
it good for classification into three categories.

For X2, 0.5 is set to capture missing values.

par(mfrow=c(1,2))

plot(seq(1,9711,length=9711),arrow$nof.MeSH.in.common,main = "Scatter plot of original number of MeSH in common", xlab = "Record No.")

plot(seq(1,9711,length=9711),X2,main = "Scatter plot of transformed number of MeSH in common (X2)", xlab = "Record No.")

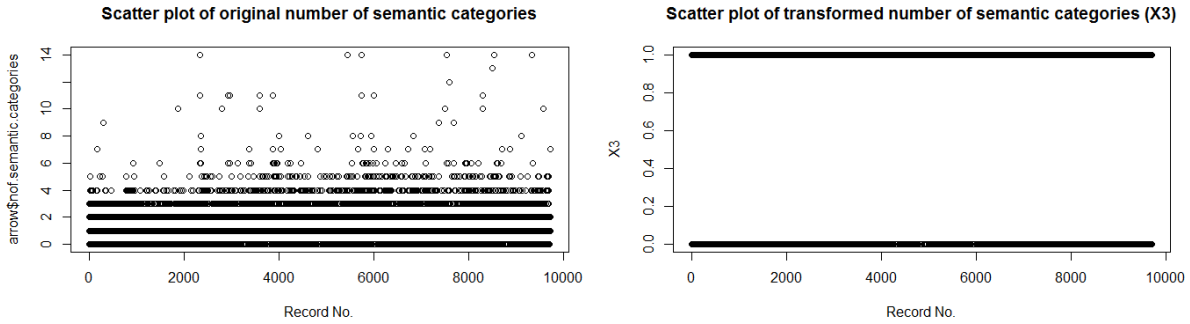


Original numbers of common MeSH mostly are relatively small, 99999 is the value for missing records.

Transformed X2 classified them into three distinct values.

For X3,

```
par(mfrow=c(1,2))

plot(seq(1,9711,length=9711),arrow$nof.semantic.categories,main = "Scatter plot of original number of semantic categories", xlab = "Record No.")

plot(seq(1,9711,length=9711),X3,main = "Scatter plot of transformed number of semantic categories (X3)", xlab = "Record No.")
```

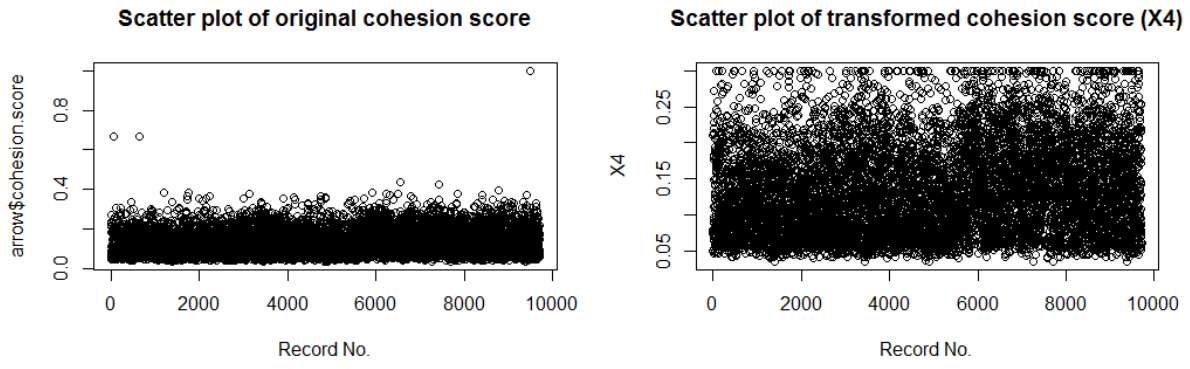


The original values of semantic categories have several different numbers.

While X3 classified them into binary values.

For X4, 0.3 is the cut off value in order to avoid nonlinearities observed at very high values.

```
par(mfrow=c(1,2))

plot(seq(1,9711,length=9711),arrow$cohesion.score,main = "Scatter plot of original cohesion score", xlab = "Record No.")

plot(seq(1,9711,length=9711),X4,main = "Scatter plot of transformed cohesion score (X4)", xlab = "Record No.")
```

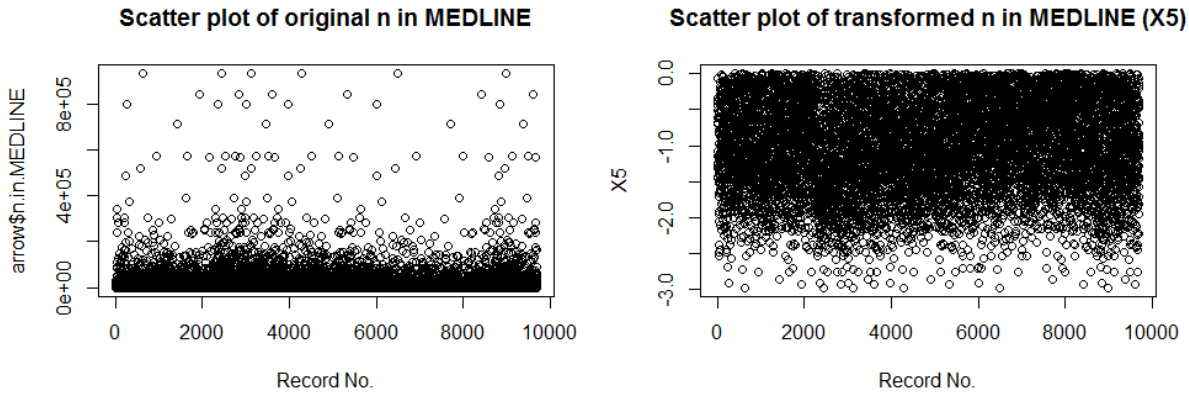


We can see that after transformation the distribution is not concentrated in a small part of the whole.

The cut off value is meaningful.

For X5, the log transformation is useful to make it easier to discover the variation.

```
par(mfrow=c(1,2))
plot(seq(1,9711,length=9711),arrow$n.in.MEDLINE, main = "Scatter plot of original n in MEDLINE", xlab = "Record No.")
plot(seq(1,9711,length=9711),X5, main = "Scatter plot of transformed n in MEDLINE (X5)", xlab = "Record No.")
```

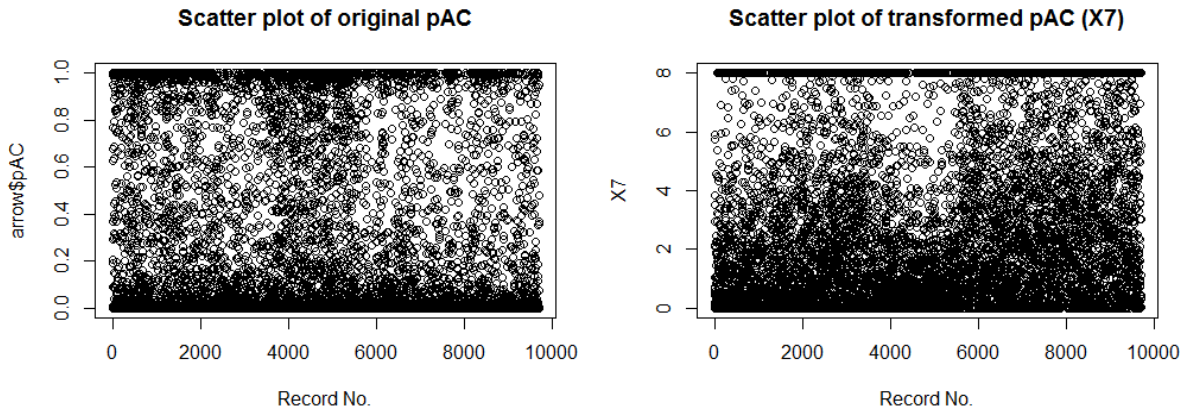


We can see that after transformation most values are spread nearly the whole range. But

before transformation most values are concentrated in a small range, which is hard for differentiation.

Similarly, for X7, log transformation is very useful.

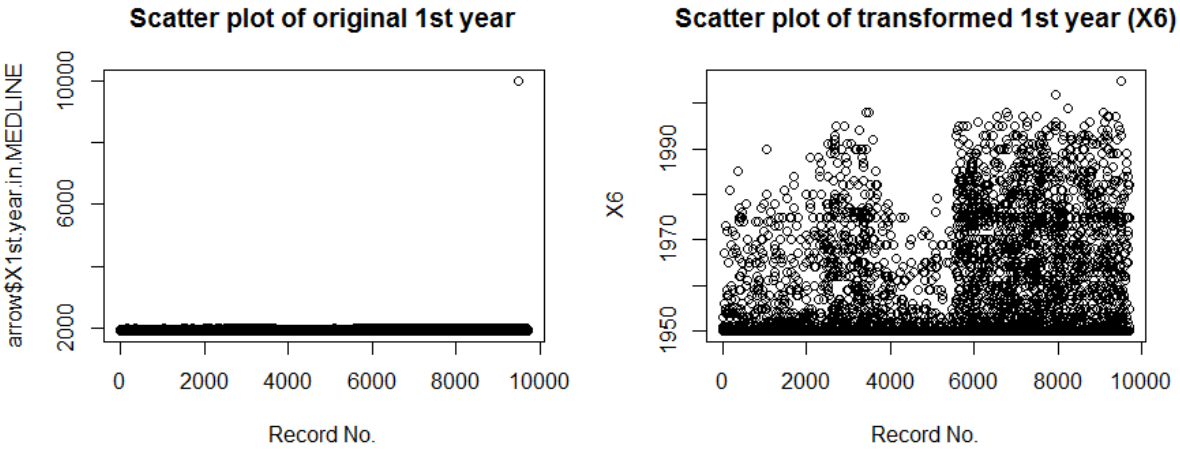
```
par(mfrow=c(1,2))
plot(seq(1,9711,length=9711),arrow$pAC, main = "Scatter plot of original pAC", xlab = "Record No.")
plot(seq(1,9711,length=9711),X7, main = "Scatter plot of transformed pAC (X7)", xlab = "Record No.")
```



After transformation it becomes more regular in distribution and the cut off value matters as well.

For X6,

```
par(mfrow=c(1,2))

plot(seq(1,9711,length=9711),arrow$X1st.year.in.MEDLINE, main = "Scatter plot of original 1st year", xlab = "Record No.")

plot(seq(1,9711,length=9711),X6, main = "Scatter plot of transformed 1st year (X6)", xlab = "Record No.")
```



The transformation obviously helps to avoid the missing value (outlier) in the original one.

missing values or outliers

Missing values are annotated by 99999 (MeSH), 0.9999 (cohesion), and 9999 (year).

```
which(arrow$cohesion.score==0.9999)
# [1] 9506

which(arrow$X1st.year.in.MEDLINE==9999)
# [1] 9506

which(arrow$nof.MeSH.in.common==99999)
length(which(arrow$nof.MeSH.in.common==99999))
# [1] 765

arrow[9506,]
```

So the 9506th record has missing values in 1st year in MEDLINE, number of MeSH and cohesion score.

There are 765 records which have missing values in number of MeSH.

For missing values in number of MeSH, X2 has already addressed it. The X2 value is set to 0.5

(median number between 0 and 1) for missing values.

For missing values in cohesion score, X4 also deleted it. Because 0.9999 is greater than 0.3 and

the X4 value is set to 0.3 for it.

For missing values in 1st year in MEDLINE, X6 also deleted it. Because 9999 is greater than 2005

and min() will choose 2005 as the value for it.

##################################

Step 4: Fit a logistic regression model and make assessments

model

arrowlm = glm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + I1 + I2 + I3 + I4 + I5 + I6, family = "binomial")

```
> summary(arrowlm)

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
```

NA as a coefficient in a regression indicates that the variable is linearly related to

the other variables. So the indicator variable I6 is a linear combination of some other variables.

```
> coef(arrowlm)
  (Intercept)           X1           X2           X3           X4           X5           X6
-86.14907404   0.73220457   0.98770170   1.31737690  13.76593549   0.58620715   0.03956808
           X7           I1           I2           I3           I4           I5           I6
  0.18873080   0.92685598   1.38271031   0.95634284   0.68351136  -1.10016252           NA
```

assess the assumptions and the statistical significance

The estimates and standard errors of the seven parameters are exactly the same as those in

Table S2. The z values in the summary function are equivalent to t-statistics in Table S2.

The values are also the same.

For the parameters of X1 to X7, the p-values are all far less than 0.001. Three stars in the

summary result mean that the significance level alpha is very small. It is a good regression model

based on the data. All the predictive features are statistically significant.

interpret parameters and the model

The coefficient parameters of X1 to X7 are all positive. So all of them are positive correlations.

Most of the parameters are less than 1. So the weights of them are not very strong.

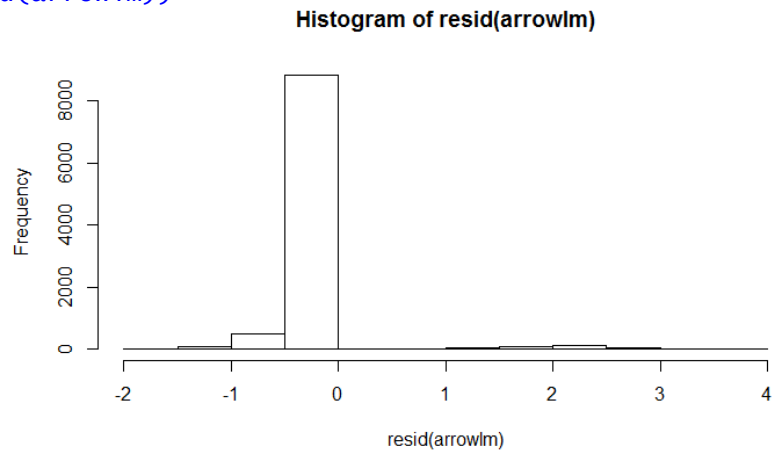
But the estimate of coefficient of X4 is much larger than others. So the value of X4

has the most significant influence on the B-term score. While the SE and z score of coefficient

of X4 are also the largest ones. So the confidence interval of the coefficient of X4 is also large.

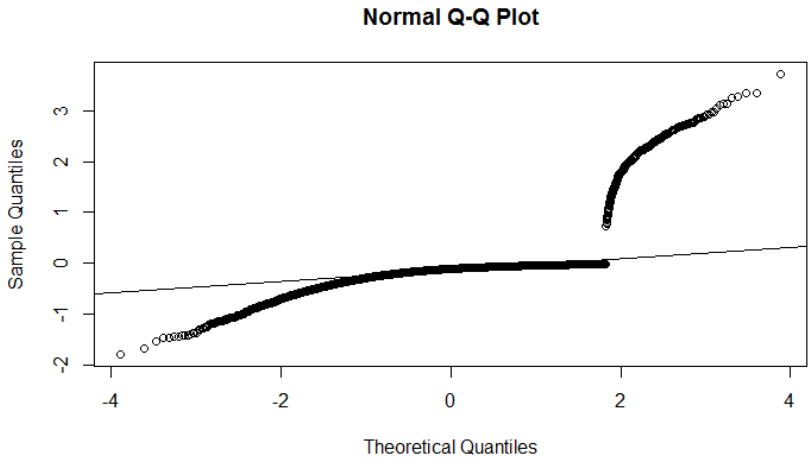
```
> summary(resid(arrowlm))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.79700 -0.21080 -0.11160 -0.10240 -0.06115  3.72700
```

```
> hist(resid(arrowlm))
```



Most residuals of the logistic regression model are close to 0. But there are some big residuals.

```
> qqnorm(resid(arrowlm))
> qqline(resid(arrowlm))
```



Most residuals are nearly normal distribution but there are some large residuals, which is not good enough.

```
> library("car", lib.loc="C:/Program Files for operation/R-3.3.1/library")
> outlierTest(arrowlm)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
     rstudent unadjusted p-value Bonferonni p
7688 3.742666         0.00018208           NA


> vif(glm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + I1 + I2 + I3 + I4 + I5, f
amily = "binomial"))
      X1       X2       X3       X4       X5       X6       X7       I1       I2       I3
1.437467 1.110059 1.024280 1.491991 1.088606 1.643758 1.327608 2.179487 1.710695 1.856529
      I4       I5
1.987519 1.925207
```

I adopted VIF to evaluate the regression model. The expression is $VIF_i = \frac{1}{1-R_i^2}$.

There is a definition: "VIF (**v**ariance **i**nflation **f**actor-a*n indicator of how much of the inflation of the standard error could be caused by collinearity*). The tolerance for a particular variable is 1 minus the R^2 that results from the regression of the other variables on that variable. The corresponding VIF is

simply 1/tolerance. If all of the variables are orthogonal to each other, in other words, completely uncorrelated with each other, both the tolerance and VIF are 1. If a variable is very closely related to another variable(s), the tolerance goes to 0, and the variance inflation gets very large."

We can see from the results that VIF of X1 to X7 are between 1 and 1.5. So these variables are basically uncorrelated with each other. So the variables are reasonable to be fit into a regression model.

```
sum(Y==1)
# [1] 326
sum(Y==0)
# [1] 9385

length(unique(arrow$B.term))
# There are 6309 unique B-terms in total.

##################################
### Step 5: Reflections
# During the replicating process, I feel that there are many details to be considered.
# First, importing data to RStudio is easy but some data cleaning need to be processed. Those actions
# are just easy commands in RStudio. Second, constructing the attributes is not an easy step. I
# debugged for several times on ifelse function. Then, making assessment of the transformation is
# a bit hard since there are lots of work regarding analysis. I took a deeper look at those expressions
# and original data. Finally, fitting a logistic model itself is easy, but interpreting
# the parameters and evaluating the model are hard. I searched on the Internet to find some
# functions for evaluation.

# When fitting the regression model, the author also added I1 to I6 as variables. It is reasonable
# to consider those six different two-node searchers. But I am not sure whether it is a must.
# If we do not include I1 to I6 as variables, the logistic regression model will be:
model = glm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, family = "binomial")
summary(model)

Call:
glm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, family = "binomial")
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3076  -0.2345  -0.1318  -0.0810   3.4233

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -52.505819   9.439127  -5.563 2.66e-08 ***
X1            0.915756   0.141361   6.478 9.29e-11 ***
X2            0.749218   0.242874   3.085  0.00204 **
X3            1.222150   0.250955   4.870 1.12e-06 ***
X4           13.378228   1.207461  11.080  < 2e-16 ***
X5            0.492119   0.111157   4.427 9.54e-06 ***
X6            0.022753   0.004859   4.683 2.83e-06 ***
X7            0.130781   0.023676   5.524 3.32e-08 ***
---
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: 2181.2  on 9703  degrees of freedom
AIC: 2197.2

Number of Fisher Scoring iterations: 8

> coef(model)
(Intercept)          X1          X2          X3          X4          X5
       X6          X7
-52.5058189   0.9157560   0.7492179   1.2221500  13.3782277   0.4921192
 0.0227526   0.1307809
```

The differences of coefficients between the two models are not significant. The coefficient of

X4 is still the largest. The weight of X1 goes up, while the weight of X2 goes down. The other

weights slightly go down a bit. Generally, the two models are basically similar. However, the

p-value of coefficient of X2 is 0.00204, which becomes larger. That means the significance level

alpha will be greater than before to satisfy the model.