math456 hw7

lgeel

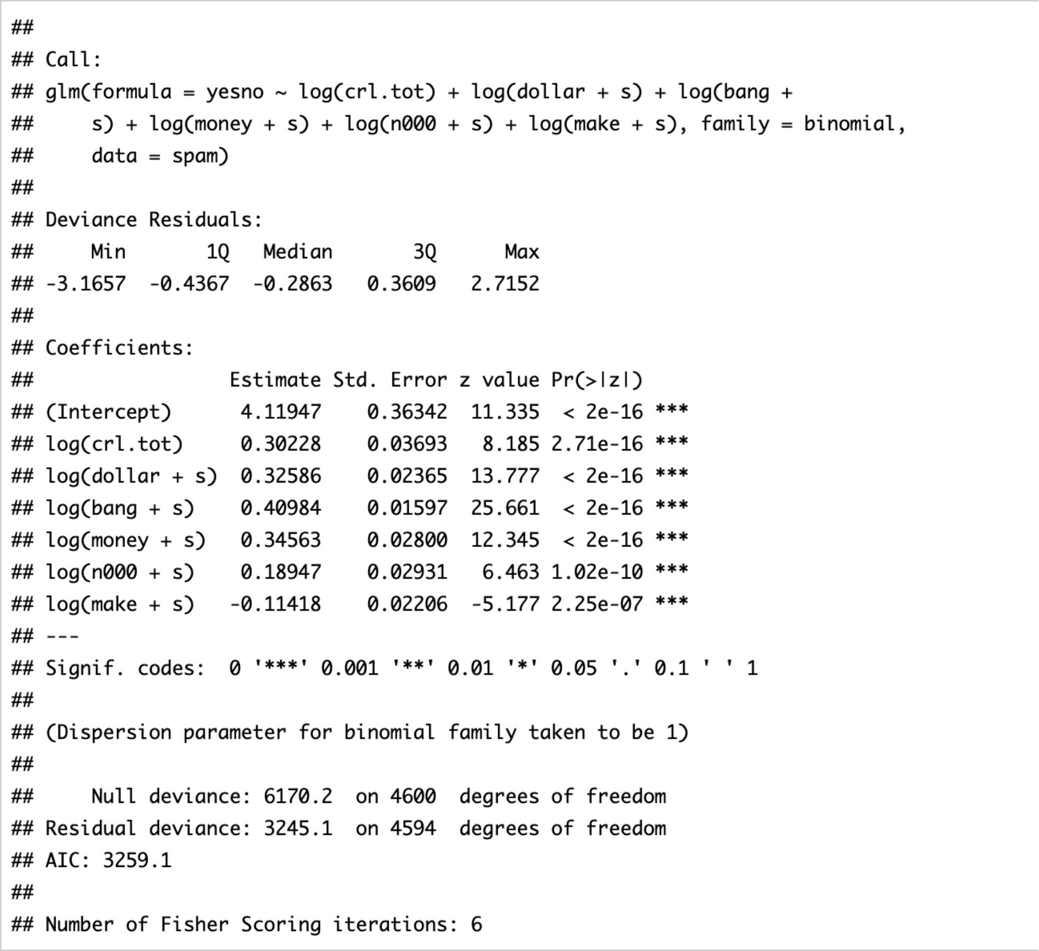
May 2022

1 Preparing data

```
import numpy as np
import pandas as pd
import string
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test _split
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
data = pd.read_csv(”spam.csv”)
data.head()
```

2 General linear model (GLM)

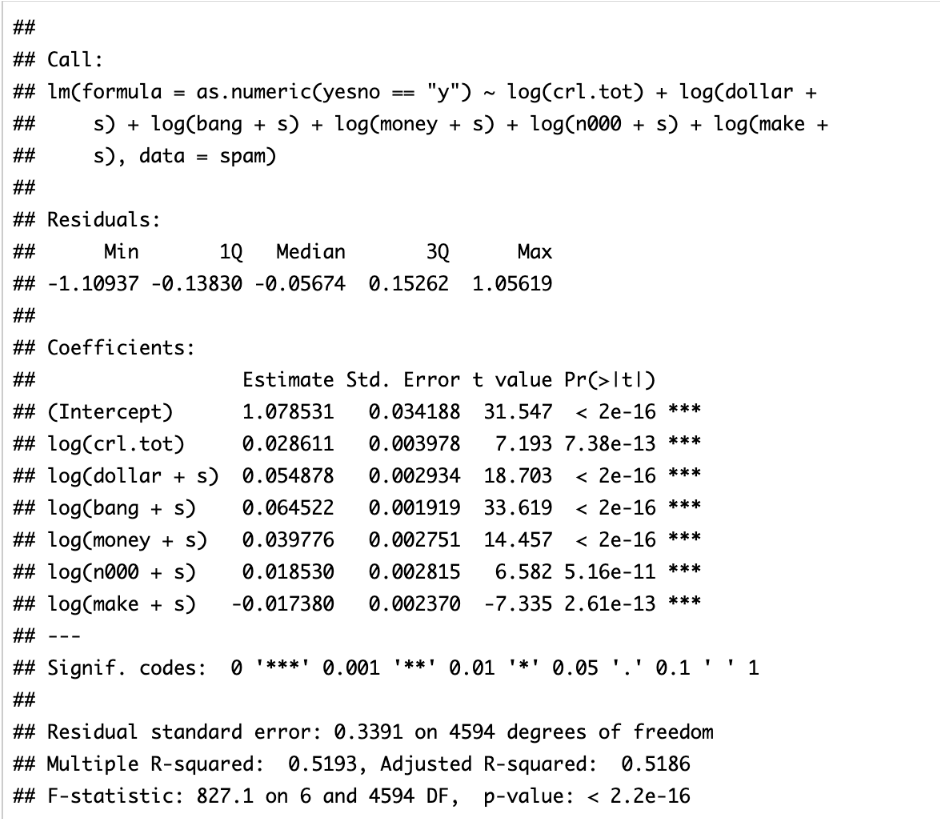
```
s = 0.001
pairs( log(crl.tot)+log(dollar+s) +log(bang+s)+log(money+s)+log(n000+s)+log(make+s)
+ yesno, data=spam, cex=.5)
spam.glm ←glm(yesno log(crl.tot) + log(dollar+s) + log(bang+s) +log(money+s)
+ log(n000+s) + log(make+s) family=binomial, data=spam)
summary(spam.glm)
```



```
##
## Call:
## glm(formula = yesno ~ log(crl.tot) + log(dollar + s) + log(bang +
##     s) + log(money + s) + log(n000 + s) + log(make + s), family = binomial,
##     data = spam)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1657  -0.4367  -0.2863   0.3609   2.7152
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.11947    0.36342  11.335  < 2e-16 ***
## log(crl.tot)     0.30228    0.03693   8.185 2.71e-16 ***
## log(dollar + s)  0.32586    0.02365  13.777  < 2e-16 ***
## log(bang + s)    0.40984    0.01597  25.661  < 2e-16 ***
## log(money + s)   0.34563    0.02800  12.345  < 2e-16 ***
## log(n000 + s)    0.18947    0.02931   6.463 1.02e-10 ***
## log(make + s)   -0.11418    0.02206  -5.177 2.25e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6170.2  on 4600  degrees of freedom
## Residual deviance: 3245.1  on 4594  degrees of freedom
## AIC: 3259.1
##
## Number of Fisher Scoring iterations: 6
```

3 Standard regression with LM

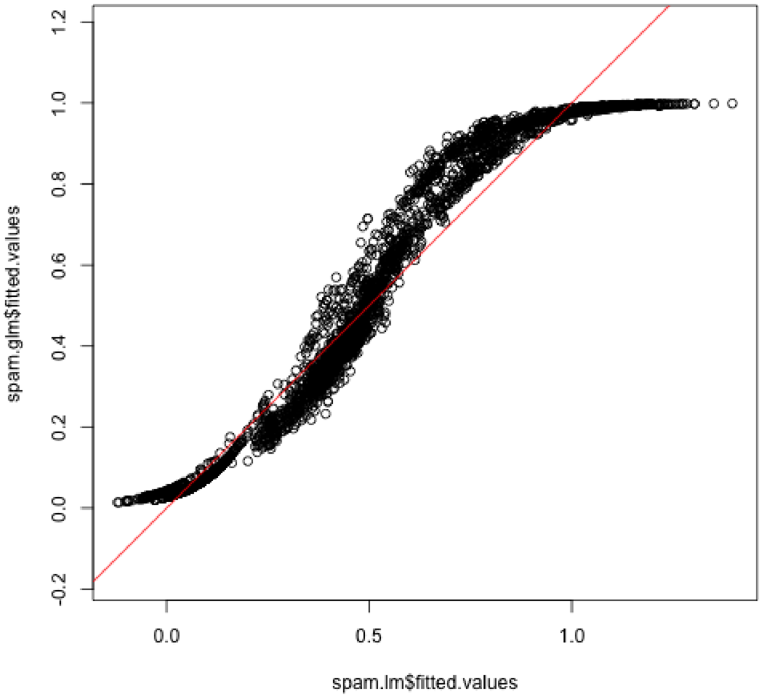
spam.lm ←lm(as.numeric(yesno==”y”) log(crl.tot) + log(dollar+s) + log(bang+s) +log(money+s) + log(n000+s) + log(make+s) ,data=spam)
$summary(spam.lm)$



```
##
## Call:
## lm(formula = as.numeric(yesno == "y") ~ log(crl.tot) + log(dollar +
##     s) + log(bang + s) + log(money + s) + log(n000 + s) + log(make +
##     s), data = spam)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10937 -0.13830 -0.05674  0.15262  1.05619
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.078531   0.034188  31.547  < 2e-16 ***
## log(crl.tot)     0.028611   0.003978   7.193 7.38e-13 ***
## log(dollar + s)  0.054878   0.002934  18.703  < 2e-16 ***
## log(bang + s)    0.064522   0.001919  33.619  < 2e-16 ***
## log(money + s)   0.039776   0.002751  14.457  < 2e-16 ***
## log(n000 + s)    0.018530   0.002815   6.582 5.16e-11 ***
## log(make + s)   -0.017380   0.002370  -7.335 2.61e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3391 on 4594 degrees of freedom
## Multiple R-squared:  0.5193, Adjusted R-squared:  0.5186
## F-statistic: 827.1 on 6 and 4594 DF,  p-value: < 2.2e-16
```

4 Comparing fitted values from the two models

par(mfrow=c(1,1))
plot(spam.lm$fitted.values, spam.glm$fitted.values,asp=1)
abline(c(0, 1), col = ”red”)


1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
spam.glm$fitted.values
0.0
0.5
1.0
spam.lm$fitted.values