Assignment 5
Shaheed Shihan
Modern Applied Statistics

6. Suppose we collect data for a group of students in a statistics class with variables $X_1 = hours$ studied, $X_2 = undergrad\ GPA$, and $Y = receive\ an\ A$. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6, \beta_1 = 0.05, \beta_2 = 1$.
(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3. 5 gets an A in the class.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$y = -6 + 40(0.05) + 3.5(1)$$
$$y = -0.5$$

$$\frac{\exp(-0.5)}{1+\exp(-0.5)} = 0.377$$

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

$$\frac{\exp(y)}{1+\exp(y)} = 0.5$$
$$y = 0$$
$$-6 + (0.05)x_1 + 3.5 = 0$$
$$x_1 = 50$$

The student has to study 50 hours to have a 50% chance to get an A.

7. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was X = 10, while the mean for those that didn't was X = 0. In addition, the variance of X for these two sets of companies was $\hat{\sigma} = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year.

$$P(X = 4|\mu = 10, \sigma^2 = 36) = \frac{\exp\frac{(-(4-10)^2}{2.36}}{\sqrt{2\pi 36}} = 0.04033$$

$$P(X = 4|\mu = 0, \sigma^2 = 36) = \frac{\exp\frac{(-(4-0)^2}{2.36}}{\sqrt{2\pi 36}} = 0.05324$$

$$P(Div = Yes|X = 4) = \frac{P(X = 4|Div).P(Div)}{P(X = 4|Div).P(Div)+P(X = 4|!\,Div).P(!Div)}$$

$$P(Div|X = 4) = \frac{0.8*0.04033}{(0.04033*0.8)+(0.05324*0.2)}$$

$$P(Div|X = 4) = 0.7519$$

There is a 75.19% chance that the company will issue a dividend this year given that the percentage profit is 4.

3.
(a)

```
> summary(weeklyMclustDA, newdata = test.X, newclass = test.Direction)
-----------------------------------------------
Gaussian finite mixture model for classification
-----------------------------------------------

MclustDA model summary:

 log.likelihood   n df      BIC
     -2129.429 985 10 -4327.785

Classes   n Model G
  Down 441     V 2
  Up   544     V 2

Training classification summary:

     Predicted
Class  Down  Up
 Down   76 365
 Up     70 474

Training error = 0.4416244

Test classification summary:

     Predicted
Class  Down Up
 Down    5 38
 Up      9 52

Test error = 0.4134615
```

I experimented with different sets of variables and the error rate was the same for the model with just lag2 and all the lags included.

```
> t
  TP        FP
1 0.5 0.04807692
```

```
(b) > summary(weeklyMclustDA, newdata = test.X, newclass = test.Direction)
----------------------------------------------------
Gaussian finite mixture model for classification
----------------------------------------------------

EDDA model summary:

 log.likelihood   n df       BIC
     -2204.237 985  3 -4429.152

Classes   n Model G
  Down 441     E 1
  Up   544     E 1

Training classification summary:

     Predicted
Class  Down  Up
  Down   22 419
  Up     20 524

Training error = 0.4456853

Test classification summary:

     Predicted
Class  Down Up
  Down    9 34
  Up      5 56

Test error = 0.375
```

```
> t3
       TP1          FP1
1 0.5384615 0.08653846
```

(c)

```
> t2
   glm      lda          qda          knn          mclustedda          mclustda
1 0.625 0.5096154    0.5865385    0.4230769    0.375               0.4134615
```

Compared to the other models both the Mclust and EDDA models have lower test errors.

4.

(a).

```
> summary(autoMclustDA, newdata = test.X2, newclass = test.mpg01)
Gaussian finite mixture model for classification
------------------------------------------------

MclustDA model summary:

 log.likelihood  n df      BIC
     -1312.097 98 39 -2803.008

Classes  n Model G
      0 47   VVE 3
      1 51   EEV 2

Training classification summary:

     Predicted
Class  0  1
   0 38  9
   1  2 49

Training error = 0.1122449

Test classification summary:

     Predicted
Class   0   1
   0 138  11
   1  33 112

Test error = 0.1496599
```

```
> data.frame(cbind(TP2,FP2))
       TP2          FP2
1 0.3809524 0.4693878
```

(b)

```
> summary(autoMclustDA, newdata = test.X2, newclass = test.mpg01)
----------------------------------------------
Gaussian finite mixture model for classification
----------------------------------------------

EDDA model summary:

 log.likelihood  n df      BIC
     -1377.171 98 18 -2836.87

Classes  n Model G
     0 47   VVV 1
     1 51   VVV 1

Training classification summary:

    Predicted
Class  0  1
   0 35 12
   1  3 48

Training error = 0.1530612

Test classification summary:

    Predicted
Class   0   1
   0 130  19
   1  15 130

Test error = 0.1156463

> data.frame(cbind(TP3,FP3))
       TP3          FP3
```

```
1 0.4421769 0.4421769
```

```
> summary(autoMclustDA, newdata = test.X2, newclass = test.mpg01)
----------------------------------------------
Gaussian finite mixture model for classification
----------------------------------------------

EDDA model summary:

 log.likelihood  n df       BIC
     -1392.842 98 17 -2863.629

Classes  n Model G
     0 47   EVV 1
     1 51   EVV 1

Training classification summary:

    Predicted
Class  0  1
   0 36 11
   1  1 50

Training error = 0.122449

Test classification summary:

    Predicted
Class   0   1
   0 131  18
   1  11 134

Test error = 0.09863946

> data.frame(cbind(TP4,FP4))
       TP4       FP4
1 0.4557823 0.4455782
```

I ran both an EDDA model and an EDDA model with “EVV” as the model name. This combination gave me the lowest error rate.

c.

```
> data.frame(cbind(glm,lda,qda,knn,automclustda,edda,EVV))
      glm       lda        qda       knn  automclustda      edda       EVV
1 0.1054422 0.0952381 0.08843537 0.1156463    0.1496599 0.09863946 0.09863946
```

Amongst all the models, qda gave me the lowest test error.

5. The paper aims at predicting the author of Ronald Reagan's radio addresses between the years of 1975 and 1979. Over a 1000 radio broadcasts were delivered during this time and while 600 of the papers were written by him, there are 312 broadcasts that have no direct evidence of who it was written by. The paper produced predictions using a variety of what they refer to as "off-the-shelf" method and other fully Bayesian generative models.

The data set they dealt with was an electronic database containing the texts of around 1032 radio addresses that Reagan had delivered during this time period. The authors draw ideas from other papers such as Augustus De Morgan's *Budget of Paradoxes* where Augustus states that it maybe possible to identify an author by the length of the words used as well as the relative frequencies.

The paper aims at distinguishing the speeches wrote by Reagan himself and those by his collaborators. They first learn the stylistic writing differences between Reagan and the others. Then EDA was used to identify a few styles that separated Reagan's literary style from the others. A fully Bayesian model was then established that allowed posterior odds of authorship to be predicted. Machine learning classification methods were then conducted on the speeches.

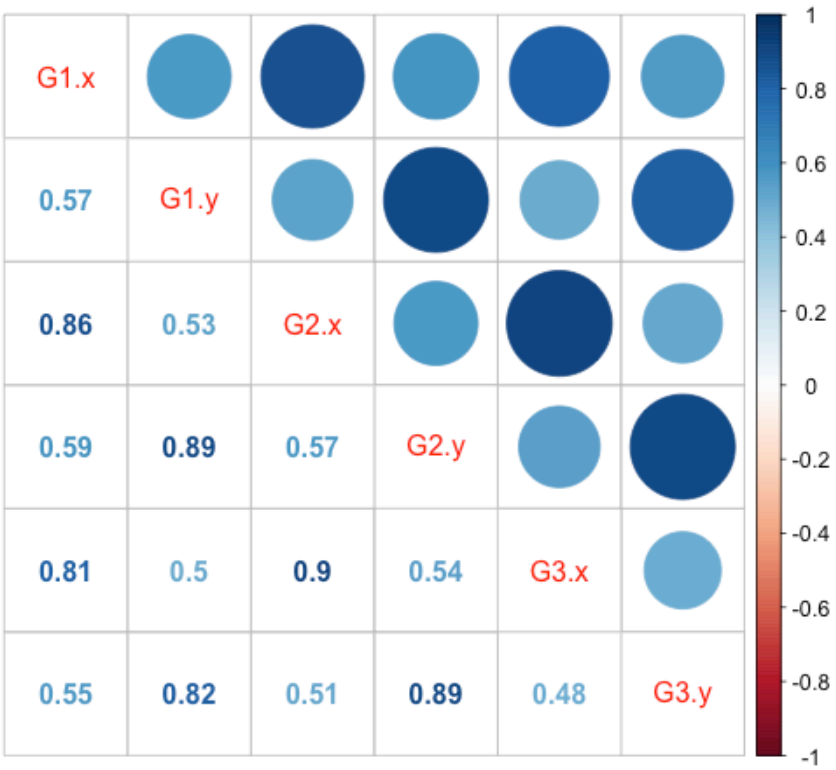
Stylistic differences were obtained by first handpicking function words (267 words) among the 3000 most commonly used words in Reagan's and Hannaford's vocabularies. A KS-statistic and t-statistic was conducted and a list of 55 discriminating words were obtained. Using the SMART text categorization system, a list of 523 words was produced that was removed from the analysis as considered not useful in the classification.

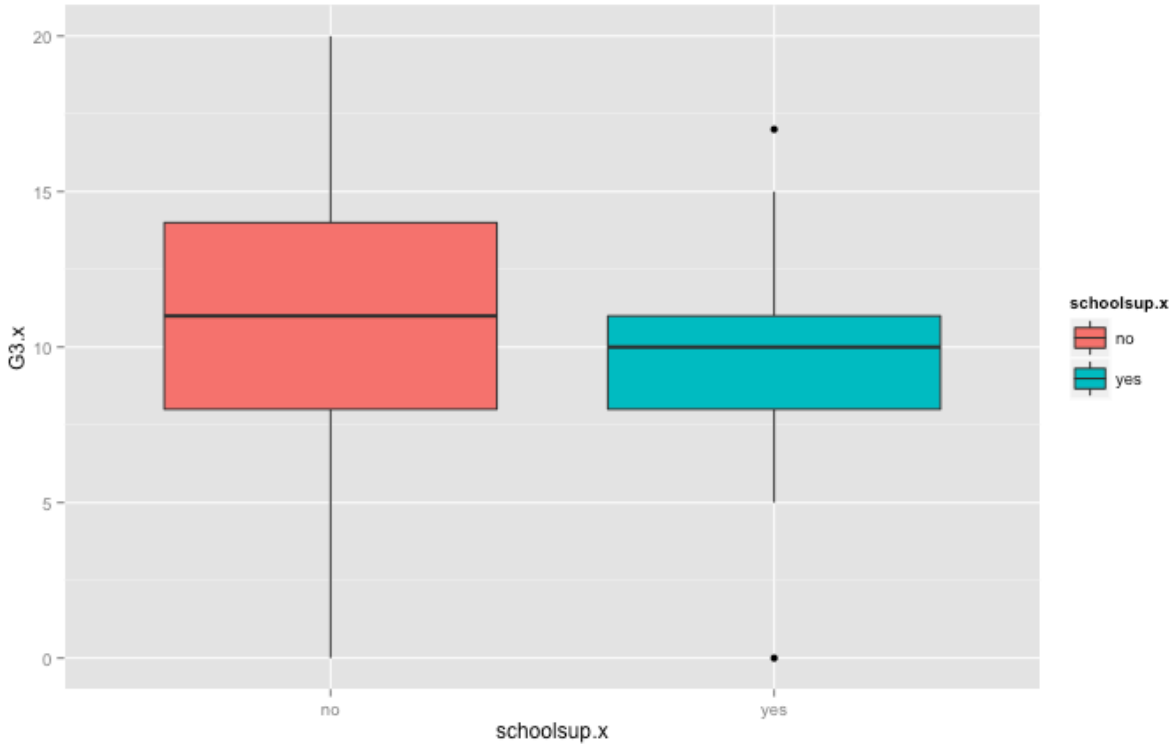
Off the shelf classifiers included linear discriminant and quadratic discriminant analysis, logistic regression, support vector machines, K-nearest neighbor, regression trees, random forests and non-parametric approaches using the maximum likelihood function.

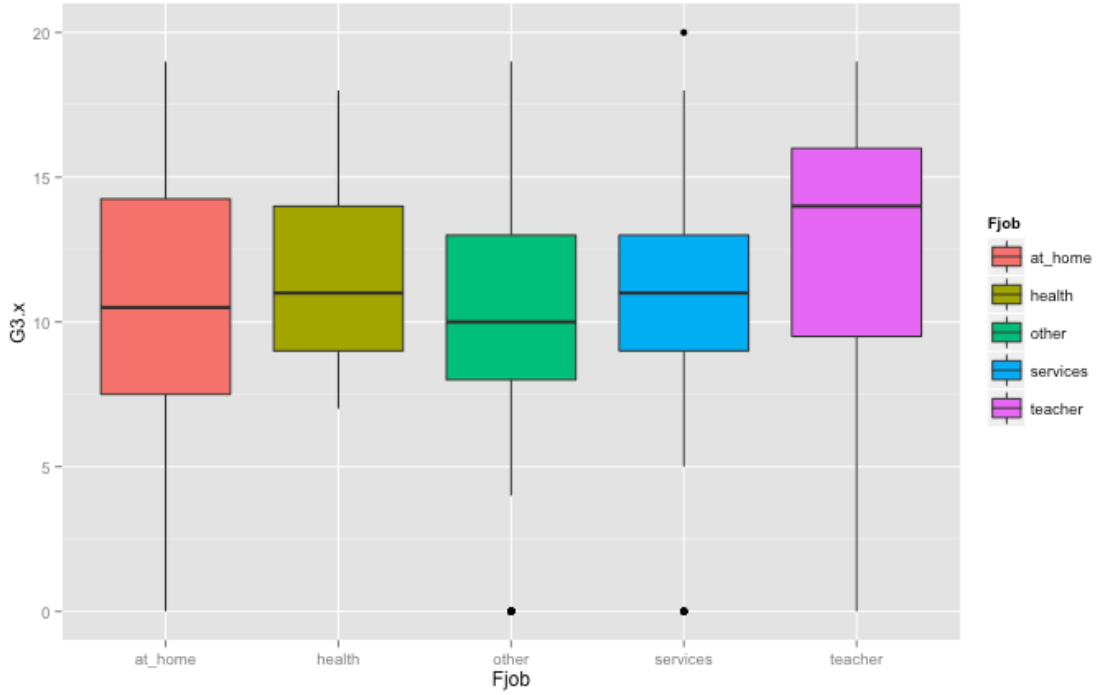
In conclusion the analysis states that the negative-binomial model is appropriate for word counts and semantic features that count data. This model also had a high predictive power in terms of predicting the authorship for the 312 "unknown" speeches.

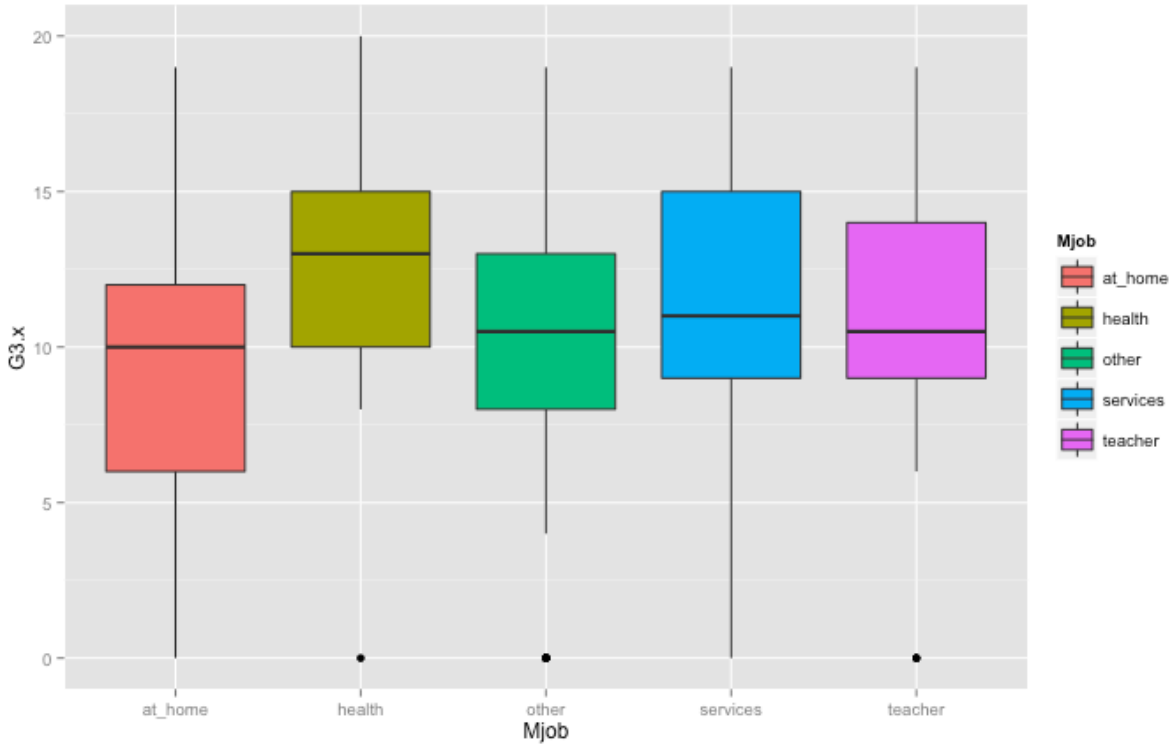
6. The sheer magnitude of the data set really threw me off. I wasn't sure where to start exploring the data or what correlations to look for. This is something I'd like more practice with if possible. The data set description does say that they are trying to predict the final grade using other variables besides the grades for the first two exams. They conducted more intensive data analysis such as Neural Networks and decision trees, which given more time, I'd definitely like to conduct.

Here are some boxplots and correlation plots that I conducted.




20
15
10
5
0
G3.x
no
yes
schoolsup.x
schoolsup.x
no
yes


20
15
10
5
0
G3.x
at_home
health
other
services
teacher
Fjob
Fjob
at_home
health
other
services
teacher


20
15
10
5
0
G3.x
at_home
health
other
services
teacher
Mjob
Mjob
at_home
health
other
services
teacher