IDS 575 : Homework 2

Team Members:

Utkarsh Ujwal Email: [uujwal2@uic.edu](mailto:uujwal2@uic.edu) UIN: 659563524

Rahul Dhanasiri Email:[rdhana4@uic.edu](mailto:rdhana4@uic.edu) UIN: 650749738

Chakradhaar Viswatmula Email:[cviswa2@uic.edu](mailto:cviswa2@uic.edu) UIN:657723905

**Question 1**

**1.a Noise, mean and variance formulation.**

**A close up of a person

Description automatically generated**

**A close up of text on a white background

Description automatically generated**

**1.b**

**Conditional Log Likelihood**

**A picture containing document, text, person

Description automatically generated**

**A close up of text on a white background

Description automatically generated**

**1.c & 1.d**

**A close up of text on a white background

Description automatically generated**

**1.e The R Code is attached with this for gradient descent algorithm on Blackboard**

**Question 2**

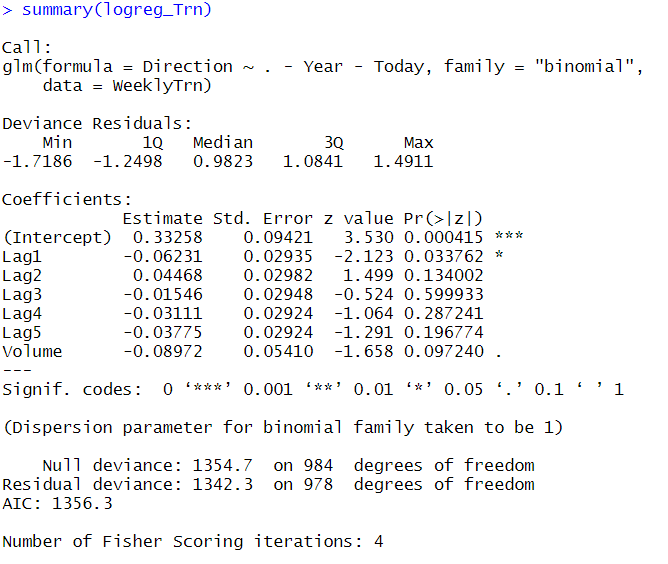
**Solution for 2.a and 2.b**

A screenshot of text

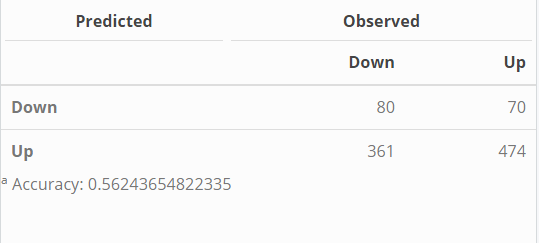
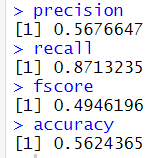
Description automatically generated

**Logistic Regression using all the variables except “Year” and “Today”:**

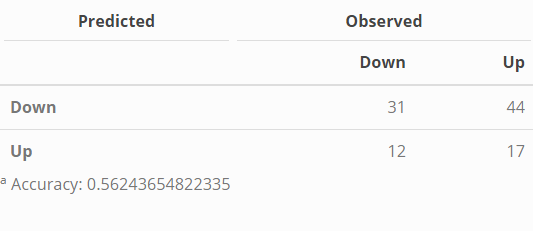
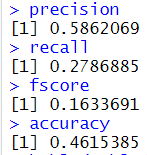
**Summary:**



**Confusion Matrix of Training Data:**

**** ****

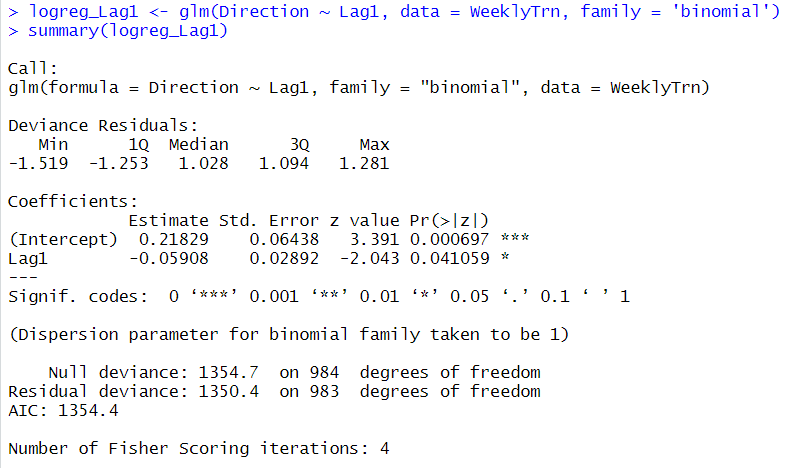
**Confusion Matrix of Test Data:**

**** ****

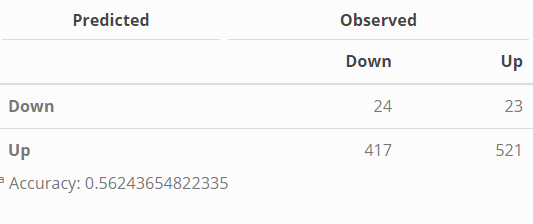
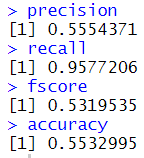
**Area under curve:** 0.5177278 **AUPR:** 0.592311

**Logistic Regression using Lag1 Variable**

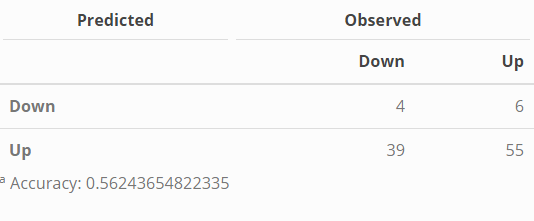
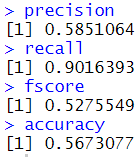
**Summary:**

****

**Confusion Matrix of Training Data:**

**** ****

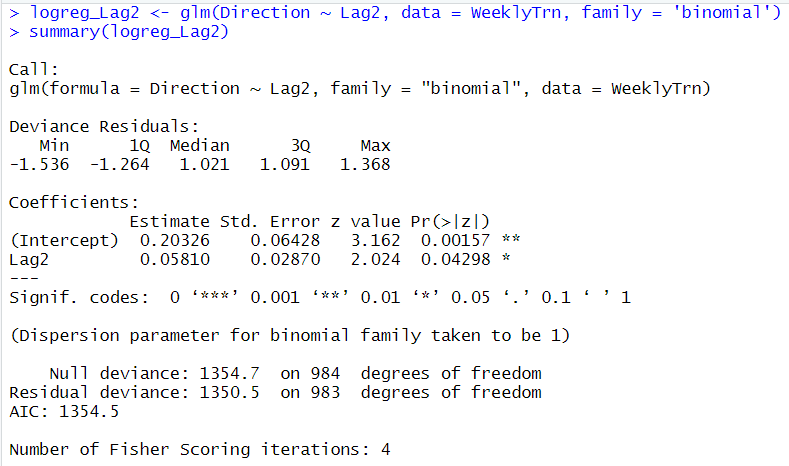
**Confusion Matrix of Test Data:**

**** ****

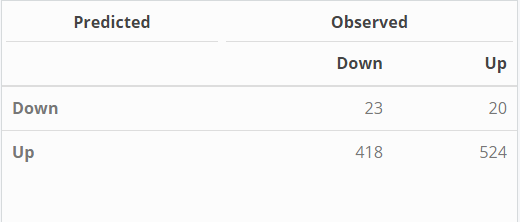
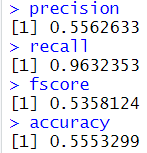
**Area under curve =** 0.4864659 AUPR: 0.5716083

**Logistic Regression using Lag2 Variable**

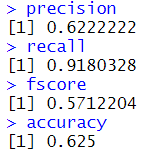
**Summary:**

****

**Confusion Matrix of Training Data:**

**** ****

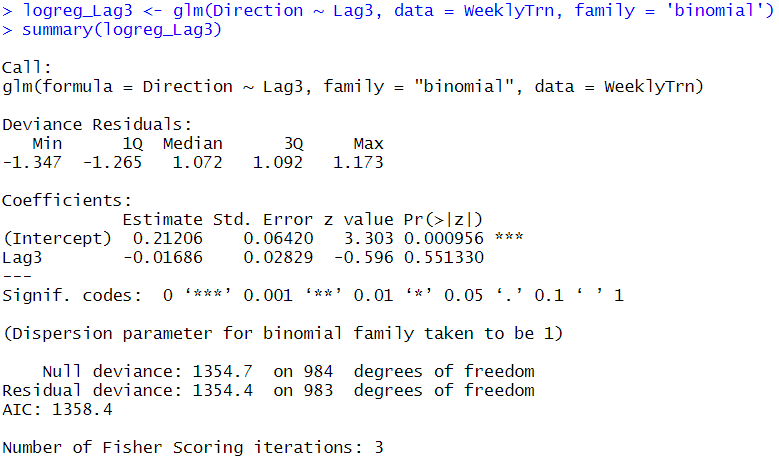
**Confusion Matrix of Test Data:**

**** ****

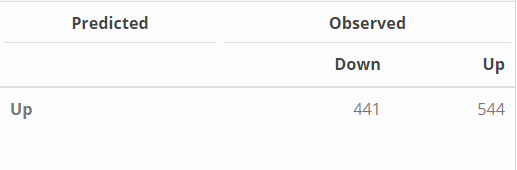
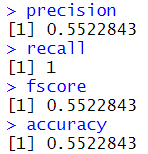
**Area under curve = 0.546321 AUPR:** 0.6421266

**Logistic Regression using Lag3 Variable**

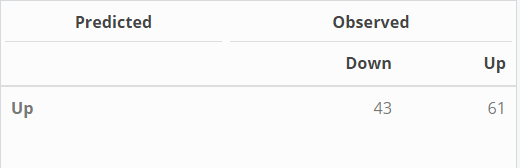
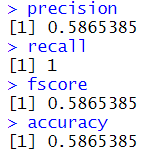
**Summary:**

****

**Confusion Matrix of Training data:**

**** ****

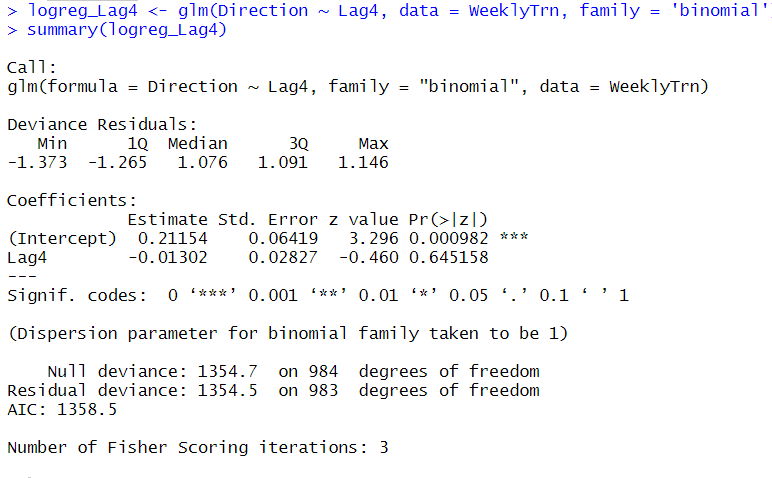
**Confusion Matrix of Test data:**

**** ****

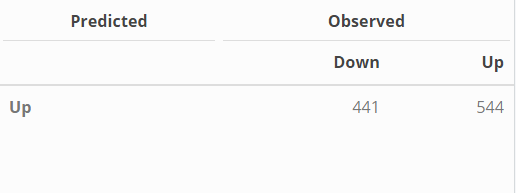
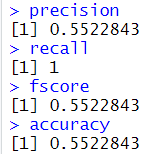
**Area under curve =** 0.5242089 **AUPR:** 0.6424658

**Logistic Regression using Lag4 Variable**

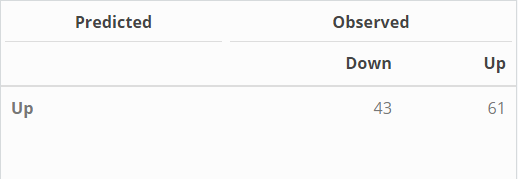
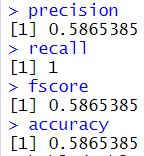
**Summary:**

****

**Confusion Matrix of Training data:**

**** ****

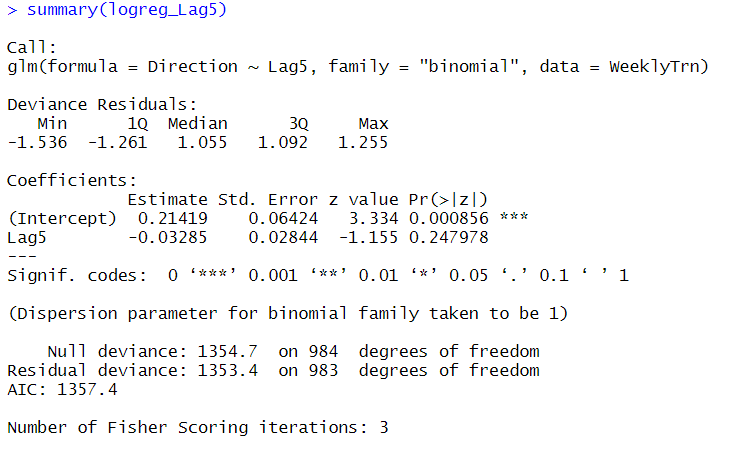
**Confusion Matrix of Test data:**

**** ****

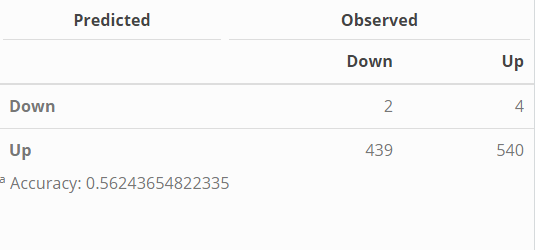
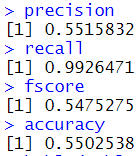
**Area under curve =** 0.5257339 AUPR: 0.6351887

**Logistic Regression using Lag5**

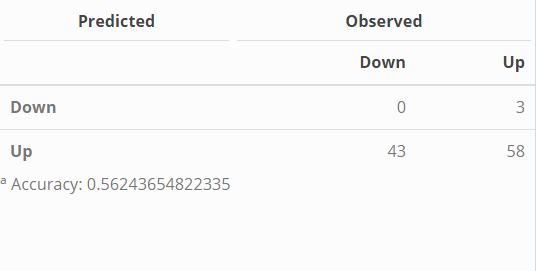
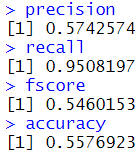
**Summary:**

****

**Confusion Matrix of Training data:**

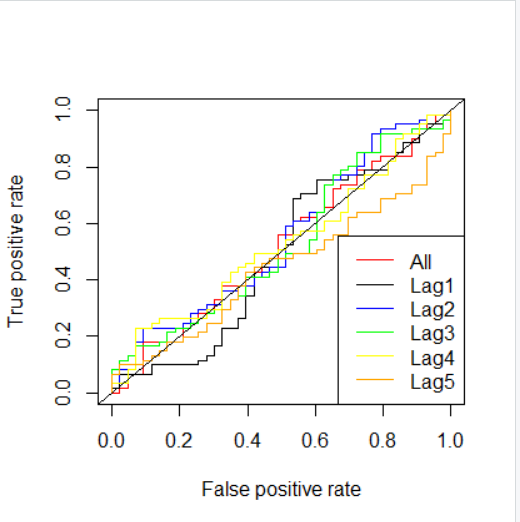
**** ****

**Confusion Matrix of Test data:**

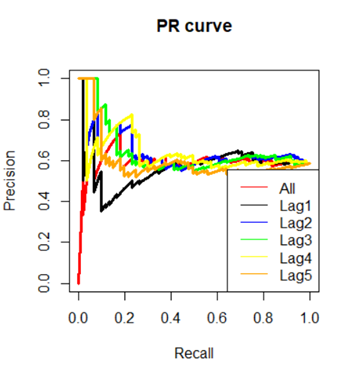
**** ****

**Area of Curve =** 0.4422417 AUPR:0.598154

* **ROC Curves for all the classifiers:**

****

**PR Curve for all classifiers:**

****

**The summary for all things answered in question 2/**

1. The probability is **0.38 (38%)** for a student to pass if he studies for 40 hours and has a GPA of 3.5
2. A student needs to study **50 hours.**
3. Confusion Matrix and Accuracy of each model is specified above.
4. The model created using variable “Lag2” is the best model as the **accuracy = 0.625 (62.5%)** and **F-score = 0.571 (57.1%).** No, the best model does not achieve the best accuracy or F-score on the training data.
5. The best model in terms of the Area Under Curve (AUC) is the model created using only Lag 2 variable **0.546321 (54.7%)**
6. The best model in terms of the Area Under Curve (AUC) is the model created using only Lag 3 variable **0.6424658 (64.24%)**
7. Based on the performance evaluation metrics like: **Precision = 62.2%**, **Recall = 91.8%,** **Accuracy = 62.5%, AUC = 54.63%, F-Score = 57.12%**

The model created using “Lag2” variable is the best model. The only statistically significant variable is “Lag2” which can be understood by the stochasticity of the data. Hence, Model created using only “Lag2” is the best model for this problem.

**Question 3**

A screenshot of a cell phone

Description automatically generated

A close up of text on a white background

Description automatically generated

A close up of text on a white background

Description automatically generated

**Question 4**

4.a

Using the equation for finding the hyperplane, yi(wT\*x+b) = 0 we get out optimal hyperplane to be x-y-1/2 =0

A close up of a map

Description automatically generated

4.b

The rules of classification are as follows:

Yes if x-y-1/2 >0

No if x-y-1/2<0

4.c

Yes the data is linearly separable. The geometric margin and support vectors are mentioned in plots above.

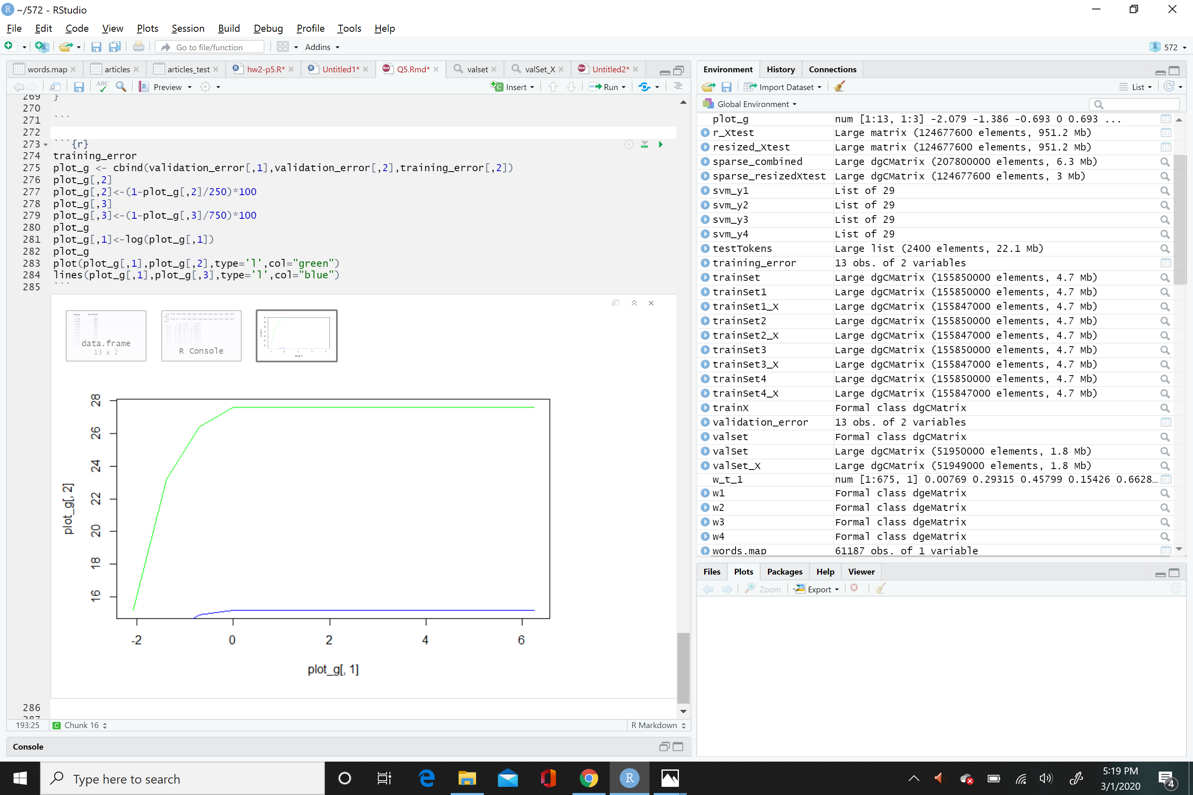
4.d

A slight perturbation of the 6th sample does not affect the optima marginal hyperplane because the 6th sample is most negative.

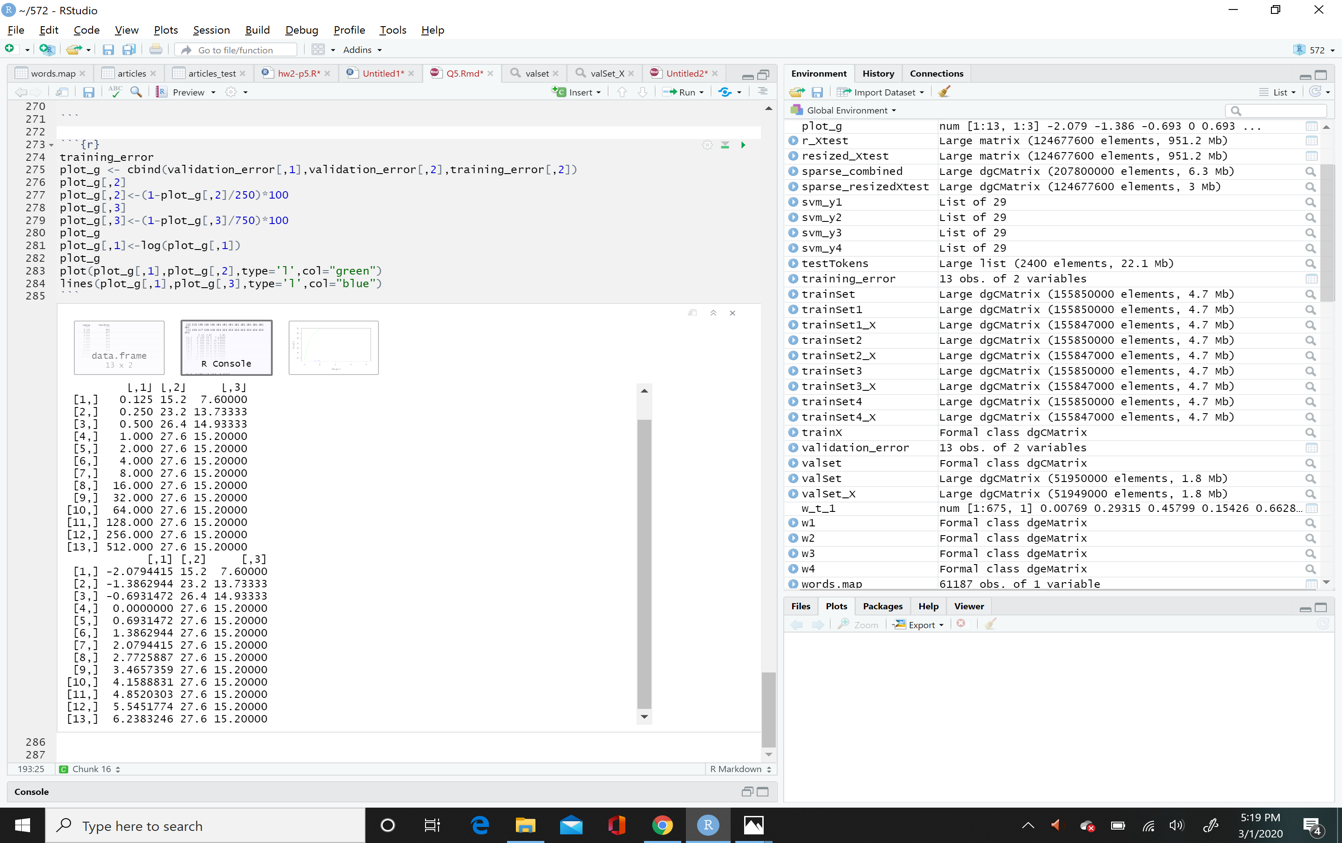
Question 5

The results for this questions are mentioned are mentioned in the screen shots below and further information has been commented in the R file.

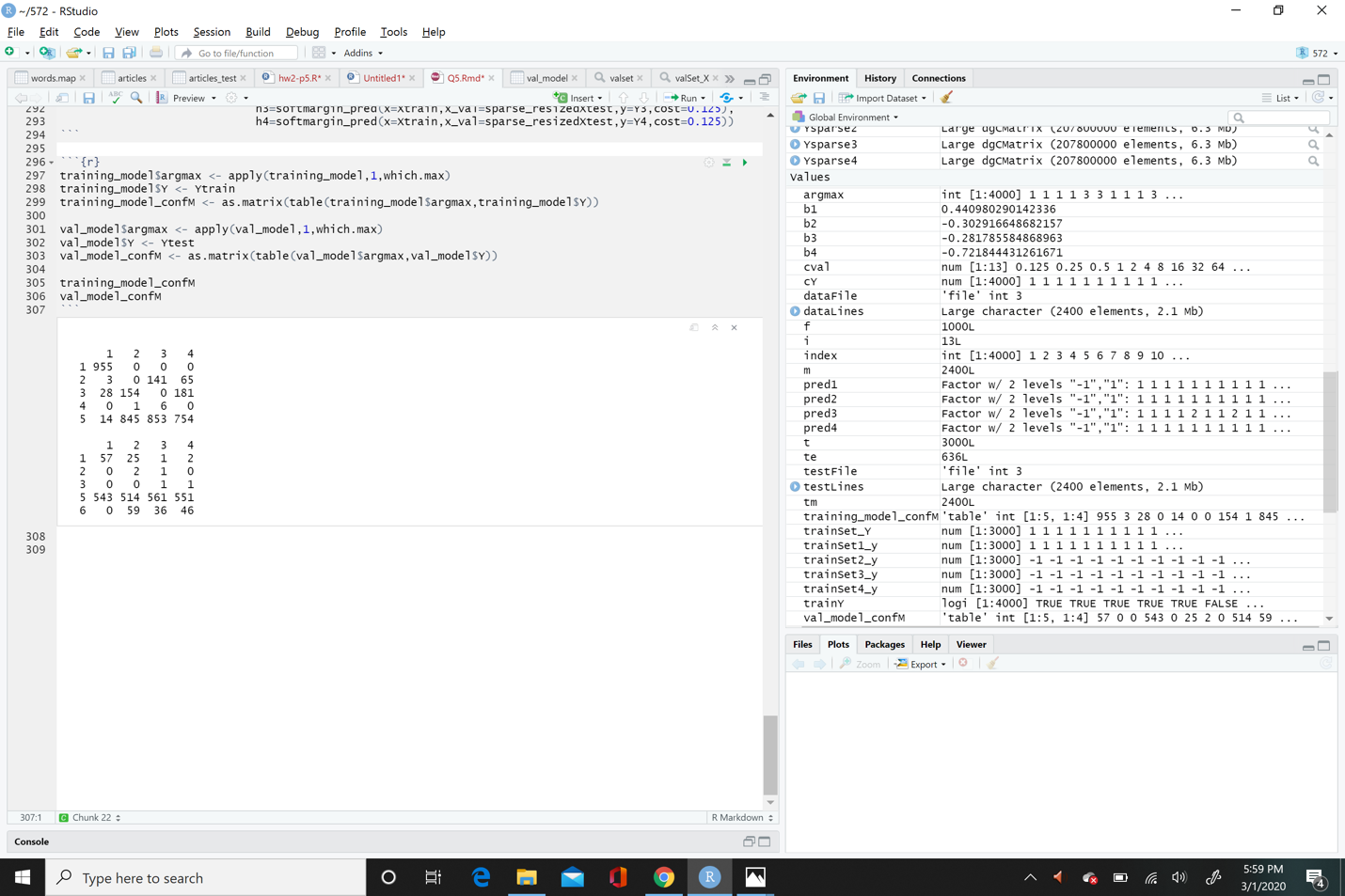
Graph showing both training and validation error together with varying C in log scale



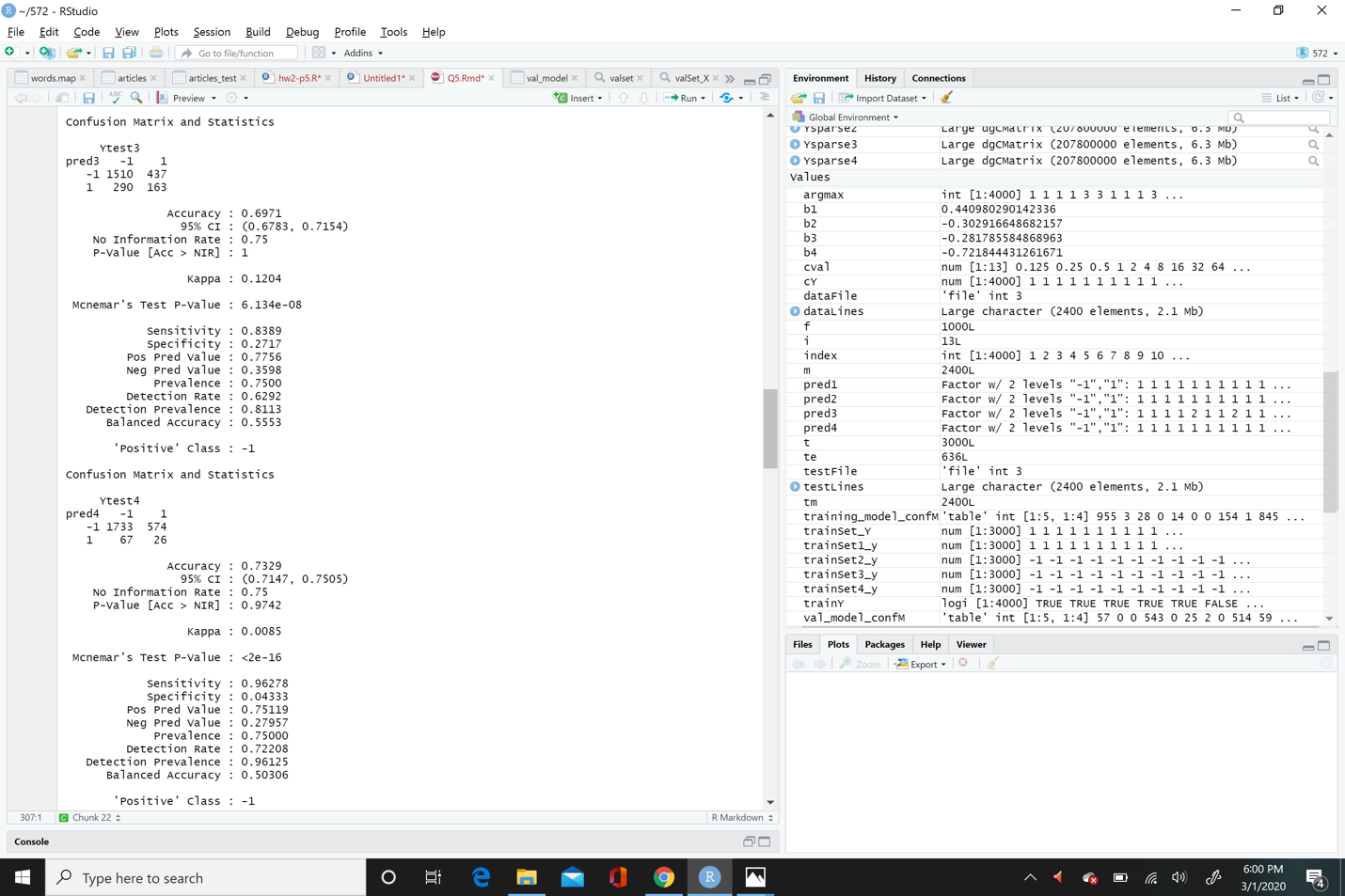
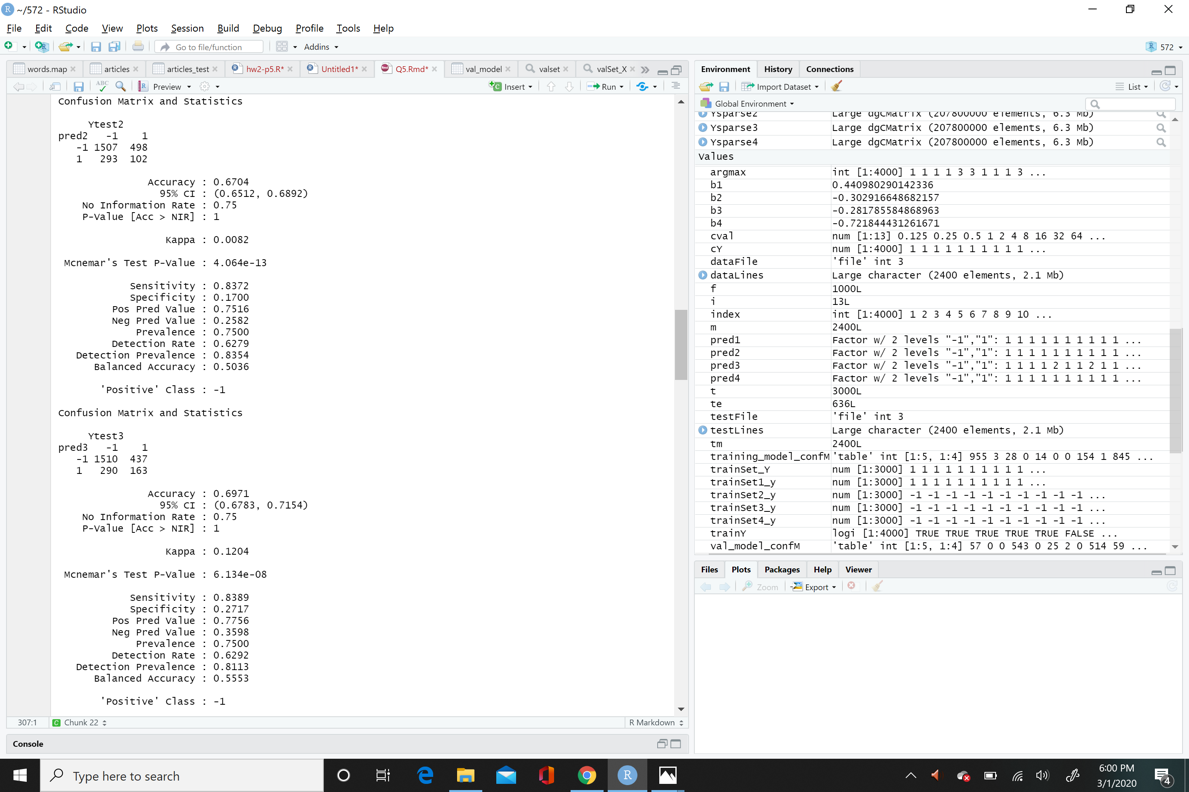
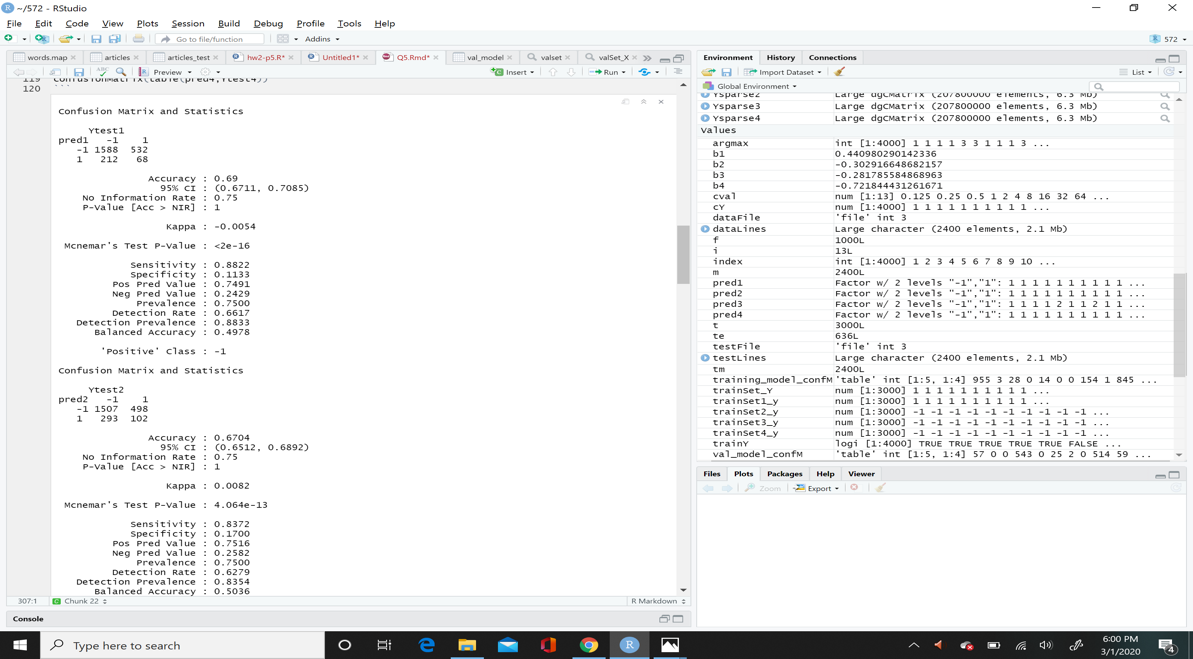
Identified best C value using the table below:



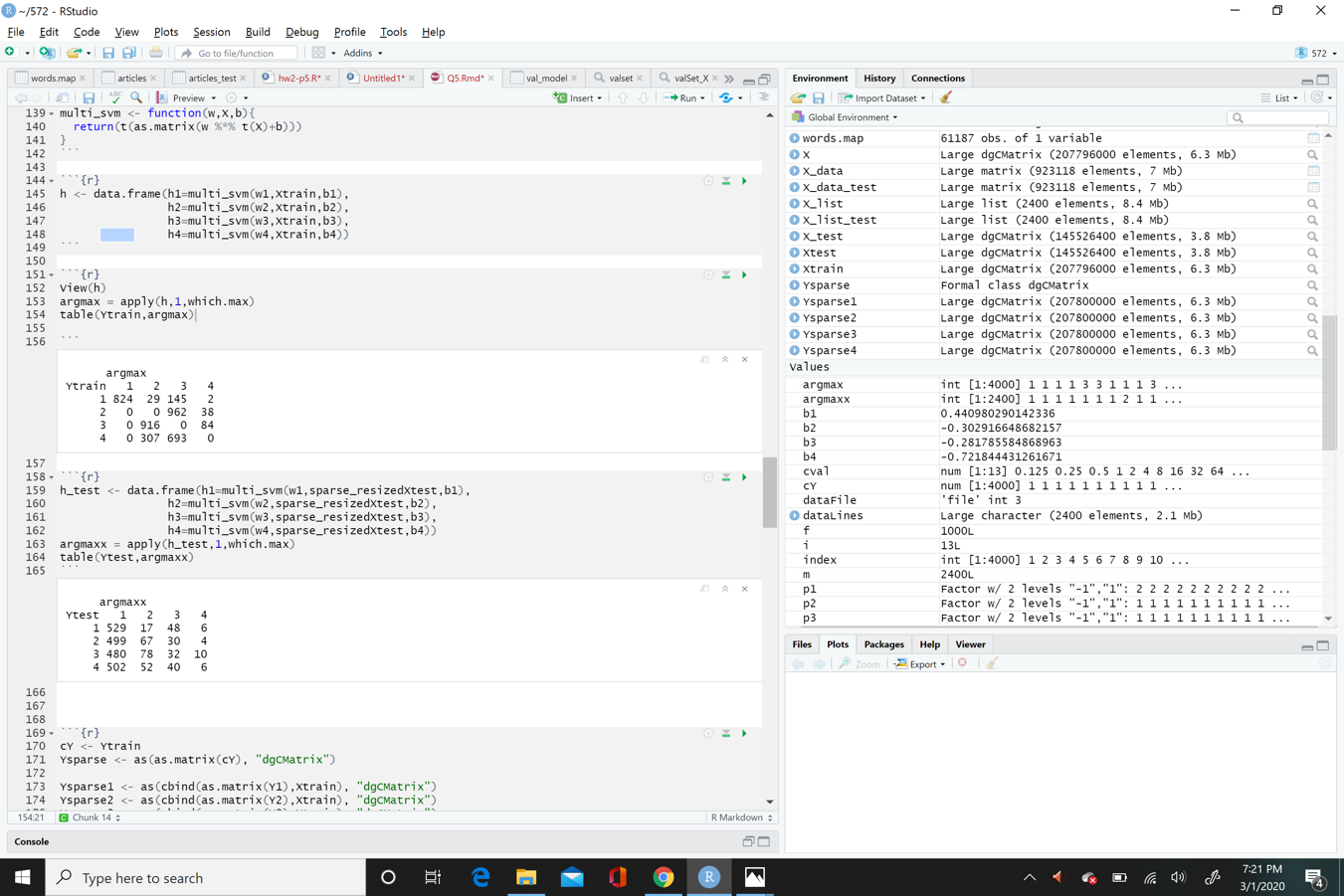
Soft margin confusion matrices (additional columns appended, please ignore).



Individual test accuracies (hard margin).



Hard margin train and test confusion matrices.



After normalizing the training and test data we got the following results:

A screenshot of a social media post

Description automatically generated

This concludes our findings for the assignment.