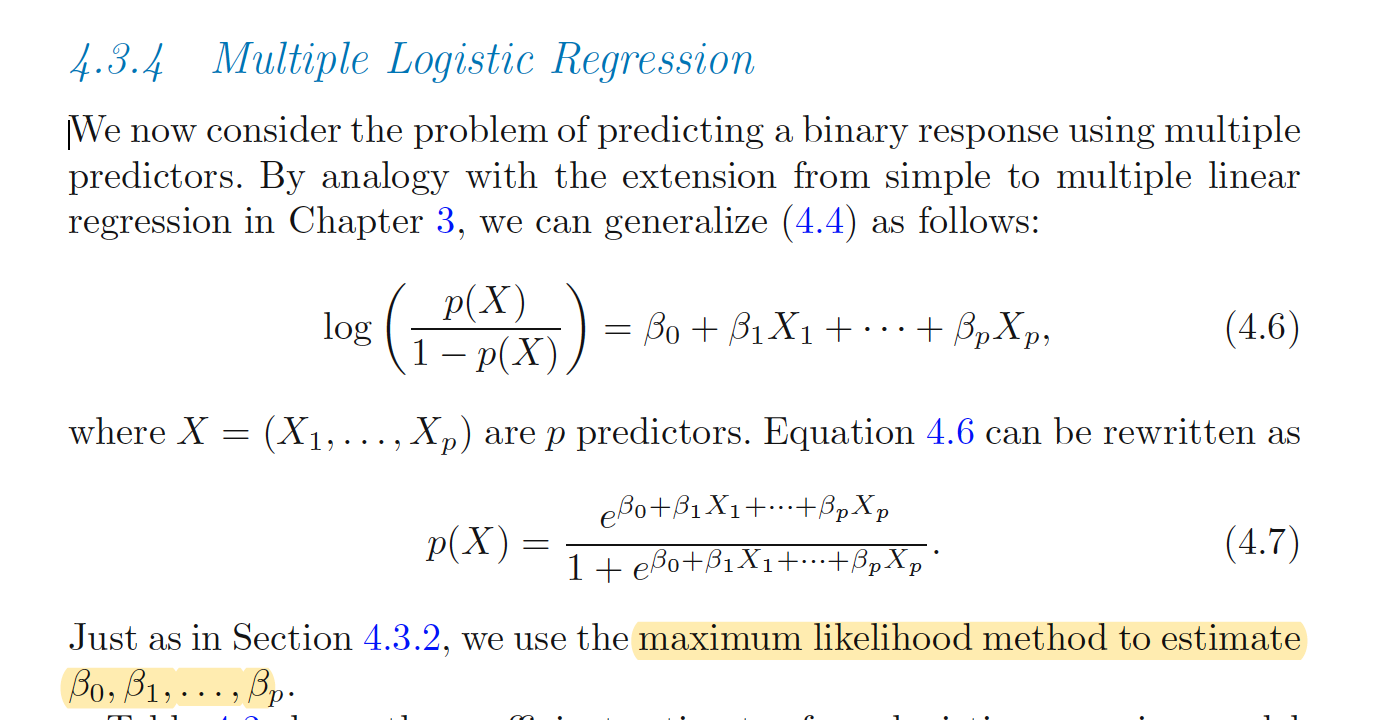
**Quiz #8**

**TOTAL POINTS 19**



1.

Question 1

Suppose we collect data for a group of students in CS 598 with variables *X*1​ = average hours studied per week, *X*2​ = undergrad GPA, and *Y* = receive an A. We fit a logistic regression with the estimated coefficients given by ***b*0​=−6, *b*1​=0.5 and *b*2​=1.**

The probability that a student who studies for **3 hours** per week (on average) and has an undergrad **GPA of 3.5** gets an A is \_\_\_\_\_\_\_\_\_\_**[a]**. (Round your answer to **2 digits after the decimal point**.)

In order to have a more than 50% chance of getting an A in Stat542, the student needs to study at least \_\_\_\_\_\_\_\_\_\_**[b]** hours per week (on average).

**[a]** = exp(b0 + (b1 \* study\_avg) + (b2 \*gpa)) / (1 + exp(b0 + (b1 \* study\_avg) + (b2 \*gpa)))

1 point



2.

Question 2

**[b]** = study\_avg = 2.5 / 0.5

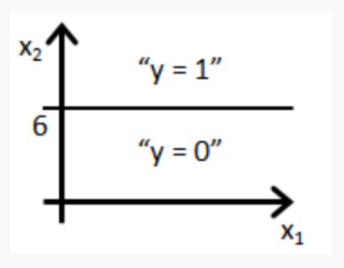
1 point



3.

Question 3

Suppose we have trained a logistic classifier with **(*b*0​+*b*1​*x*1​+*b*2​*x*2​)** where the feature vector ***x*=(*x*1​,*x*2​)** is two-dimensional and ***b*0​=6, *b*1​=0, and *b*2​=−1**. Suppose we classify data points to be 1 if the estimated probability (from the logistic model) is bigger than **0.5, and 0** otherwise. What is the decision boundary of our classifier? (x1 does not seem to be important since coeff is zero. B0 and b2 and hence X2 seem to be important) , ref: <https://github.com/zhanwen/machine-learning/blob/master/coursera/week3/logistic_regression.md> - see last diagram



1 point



Decision boundary is the line *x*2​=6: data points with *x*2​>6 will be classified as *Y*=0 and data points with *x*2 ​≤ 6 will be classified as *Y*=1.

(Refer to above github. if you notice the coeff’s since the negatives the diagram will be flipped I guess, y = 1 for x2<=6 makes sense.)



Decision boundary is the line *x*1​=6: data points with *x*1​>6 will be classified as *Y*=1 and data points with *x*1​≤6 will be classified as *Y*=0.



Decision boundary is the line *x*2​=6: data points with *x*2​>6 will be classified as *Y*=1 and data points with *x*2​≤6 will be classified as *Y*=0. (or this)



Decision boundary is the line *x*1​=6: data points with *x*1​>6 will be classified as *Y*=0 and data points with *x*1​≤6 will be classified as *Y*=1.

4.

Question 4

We estimate the logistic regression coefficient vector \beta*β* by minimizing J(\beta) = *J*(*β*)= (-log) likelihood function. Which **three** of the following statements about this cost function J(\beta)*J*(*β*) are correct?

1 point



The MLE of \beta*β*, i.e., the minimizer of J(\beta)*J*(*β*), may not exist.



The Newton-Raphson algorithm, which we use to find the minimizer of J(\beta)*J*(*β*), could get stuck at a local minimum, even if the global minimum exists.



The cost function J(\beta)*J*(*β*) for logistic regression is only positive at some \beta*β* values, e.g., at the MLE of \beta*β*.



The cost function J(\beta)*J*(*β*) for logistic regression is always non-negative.



The cost function J(\beta)*J*(*β*) for logistic regression is convex, so any local minimum is a global minimum.

5.

Question 5

**The remaining questions are related.**

Fit a logistic regression model on the "Caravan" data set from the R package "ISLR". This data set, also analyzed in Sec 4.6.6 of ISLR, has 85 predictors and the response variable is "Purchase" that is equal to "Yes" or "No".

We use the first 1000 obs as the test data and the remaining as the training data. In the test data, there are **941 "No" and 59 "Yes**". For each of the approaches below, report the number of mis-classified samples among the 941 "No" and the number of mis-classified samples among 59 "Yes", if we use 0.25 as the predicted probability cut-off. Also use the R package "pROC" to report the corresponding AUC. For the definition of AUC and ROC, read pp146-149 of ISLR.

**Fit a logistic regression model using all 85 predictors**, and obtain the predicted probabilities on the test data.

* If we use 0.25 as the probability cut-off, we misclassify \_\_\_22\_\_\_\_\_**[a1]** (an integer) samples among 941 "No" and misclassifty **\_\_\_48\_\_\_\_\_[b1]** (an integer) samples among 59 "Yes".
* The AUC for this classifier is \_\_0.741\_\_\_\_**[c1]** (round to 3 digits after the decimal point).

**Apply forward variable selection using AIC**. Use the selected model to obtain the predicted probabilities on the test data.

* We use a model with \_\_\_~~11~~\_17\_\_**[d2]** (a non-negative integer) non-intercept predictors.
* If we use 0.25 as the probability cut-off, we misclassify \_\_16\_\_\_\_ **[a2]** (an integer) samples among 941 "No" and misclassifty \_\_51\_\_\_\_ **[b2]** (an integer) samples among 59 "Yes".
* The AUC for this classifier is \_\_0.735\_\_\_\_ **[c2]** (round to 3 digits after the decimal point).

**Apply forward variable selection using BIC**. Use the selected model to obtain the predicted probabilities on the test data.

* We use a model with ~~\_\_\_4~~\_6\_\_ **[d3]** (a non-negative integer) non-intercept predictors.
* If we use 0.25 as the probability cut-off, we misclassify \_\_\_8\_\_\_ **[a3]** (an integer) samples among 941 "No" and misclassifty \_\_ 58 \_\_\_\_ **[b3]** (an integer) samples among 59 "Yes".
* The AUC for this classifier is \_\_\_0.741\_\_\_ **[c3]** (round to 3 digits after the decimal point).

**Use L1 penalty to select a subset of the predictors**. Use the glmnet package and set lambda = 0.004, and use the default options such as **standardize = TRUE**, **intercept=TRUE**. Use the selected model to obtain the predicted probabilities on the test data.

* We use a model with \_\_~~1~~\_0\_\_\_ **[d4]** (a non-negative integer) non-intercept predictors.
* If we use 0.25 as the probability cut-off, we misclassify \_\_6\_\_\_\_ **[a4]** (an integer) samples among 941 "No" and misclassifty \_\_59\_\_\_\_ **[b4]** (an integer) samples among 59 "Yes".
* The AUC for this classifier is \_\_\_0.746\_\_\_ **[c4]** (round to 3 digits after the decimal point).

**[a1]**=\_\_\_\_\_\_\_\_

1 pointm



6.

Question 6

**[b1]**=\_\_\_\_\_\_\_\_

1 point



7.

Question 7

**[c1]**=\_\_\_\_\_\_\_

1 point



8.

Question 8

**[d2]**=\_\_\_\_\_\_\_

1 point



9.

Question 9

**[a2]**=\_\_\_\_\_\_\_

1 point



10.

Question 10

**[b2]**=\_\_\_\_\_\_

1 point



11.

Question 11

**[c2]**=\_\_\_\_\_\_\_

1 point



12.

Question 12

**[d3]**=\_\_\_\_\_\_\_

1 point



13.

Question 13

**[a3]**=\_\_\_\_\_\_\_

1 point



14.

Question 14

**[b3]**=\_\_\_\_\_\_\_

1 point



15.

Question 15

**[c3]**=\_\_\_\_\_\_\_

1 point



16.

Question 16

**[d4]**=\_\_\_\_\_\_\_

1 point



17.

Question 17

**[a4]**=\_\_\_\_\_\_\_

1 point



18.

Question 18

**[b4]**=\_\_\_\_\_\_\_

1 point



19.

Question 19

**[c4]**=\_\_\_\_\_\_\_

1 point



