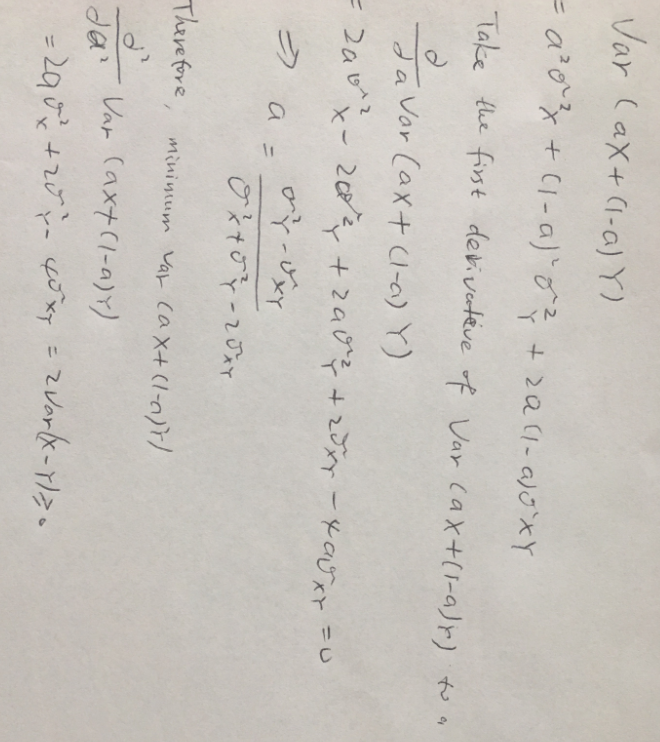
**HOME WORK THREE Qian Yu**

**ST635 Intermediate Statistical Modeling for Business Fall 2017**

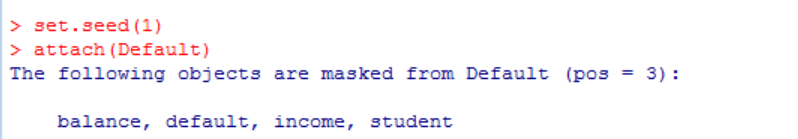
Exercise 5.4 - Problem 1, Problem 6, Problem 8

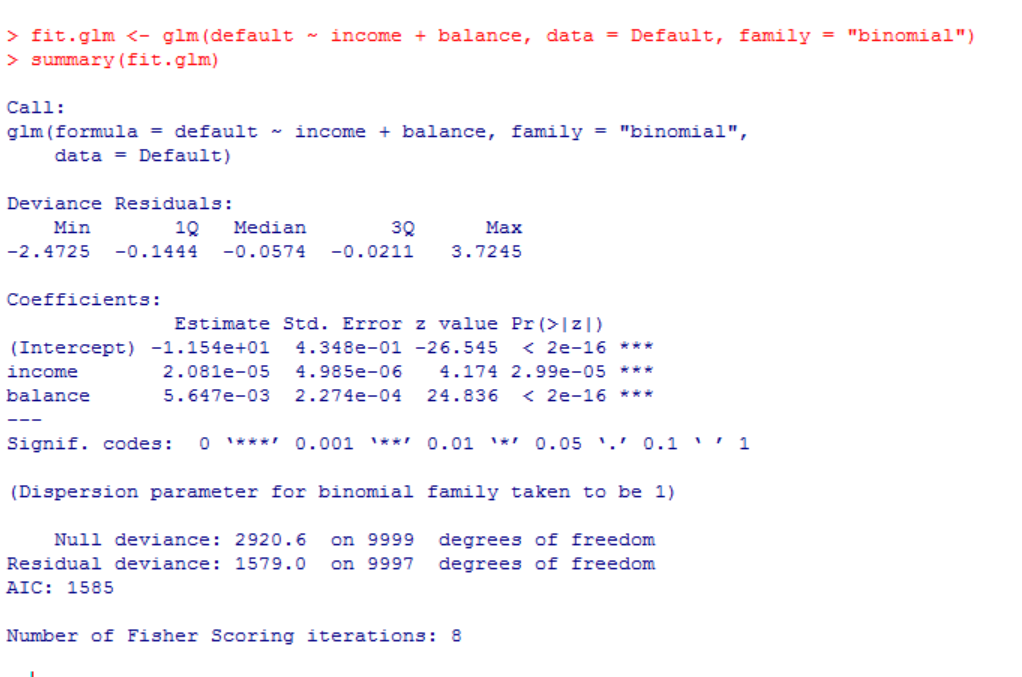
1. Using basic statistical properties of the variance, as well as single variable calculus, derive (5.6). In other words, prove that *α* given by (5.6) does indeed minimize Var(*αX* + (1 *− α*)*Y* ).



6.

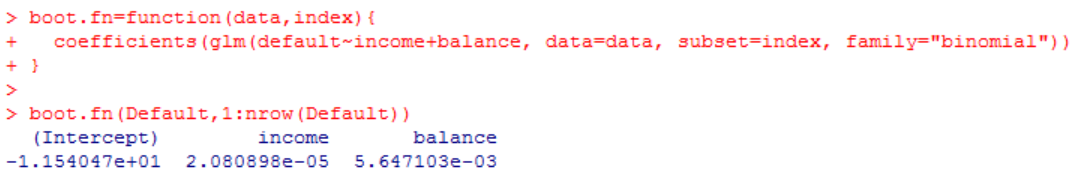
(a) Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.



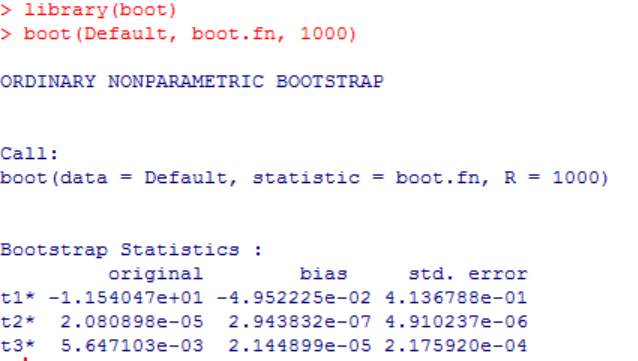


Therefore, the standard errors for the coefficients β0 (intercept), β1(income) and β2(balance) are 4.348e-01, 4.985e-06, 2.274e-04

(b) Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.



(c) Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.



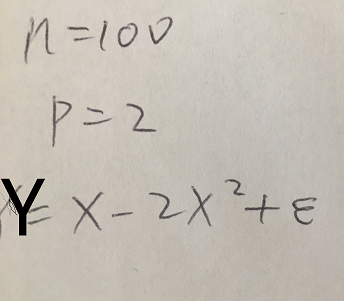
(d) Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.

*The estimated standard errors from the two methods are close to each other.*

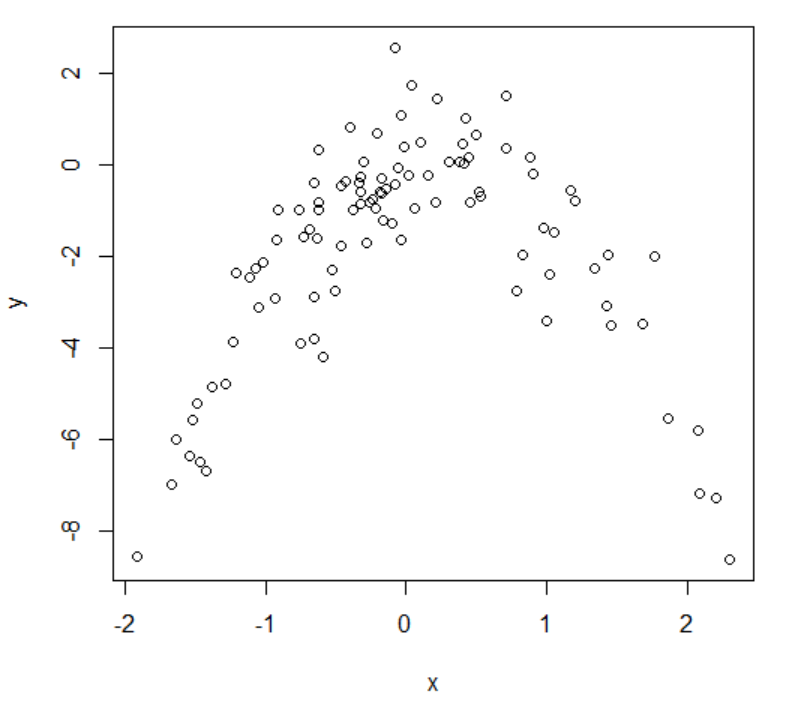
8.

(a) Generate a simulated data set as follows:

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.



(b) Create a scatterplot of X against Y . Comment on what you find.

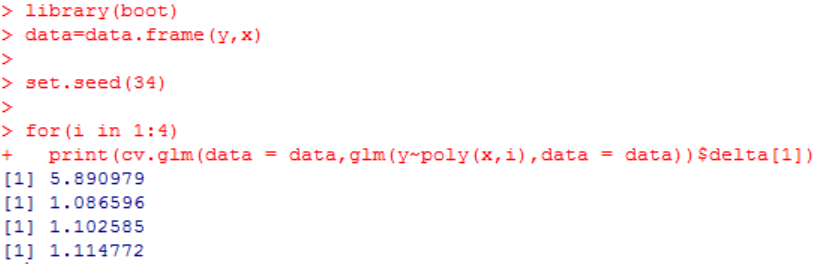


*Apply: plot(x,y)*

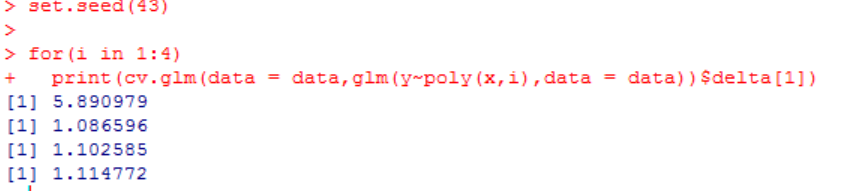
*Conclusion: There is a non-linear relationship from the parabola shape of the curve.*

(c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:

Note you may find it helpful to use the data.frame() function to create a single data set containing both X and Y .



(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

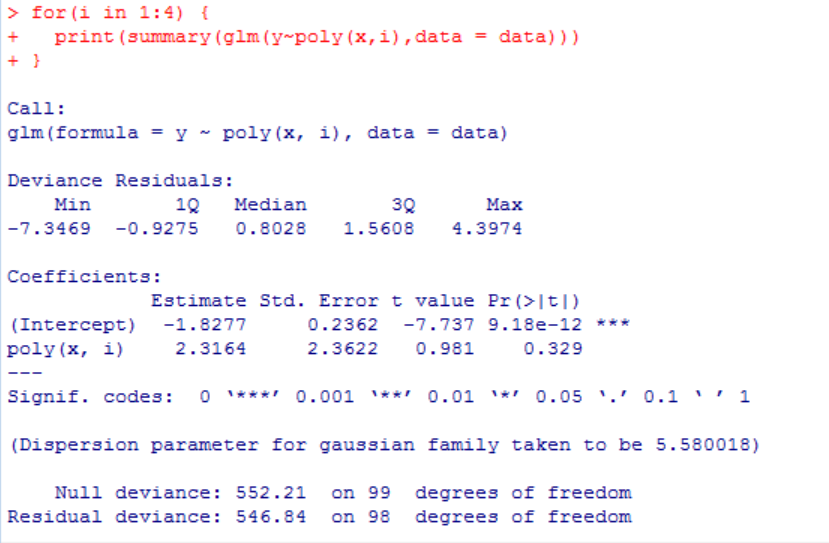


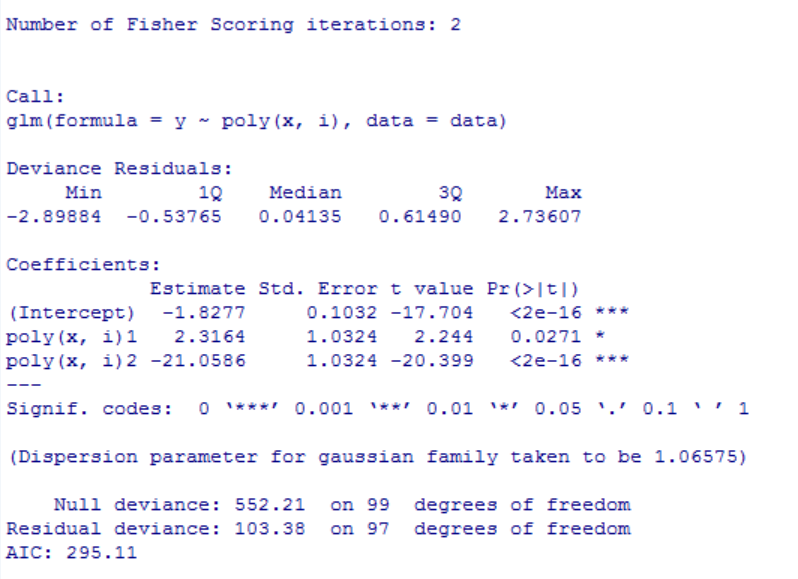
Because there is no sampling involved in LOOCV and the model is trained with the same observations for each cross validation test; therefore, the result is unchanged.

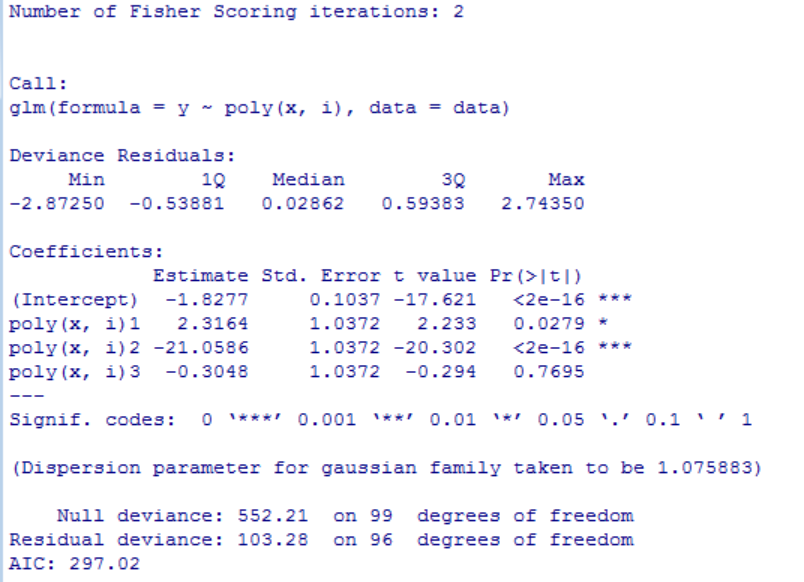
(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

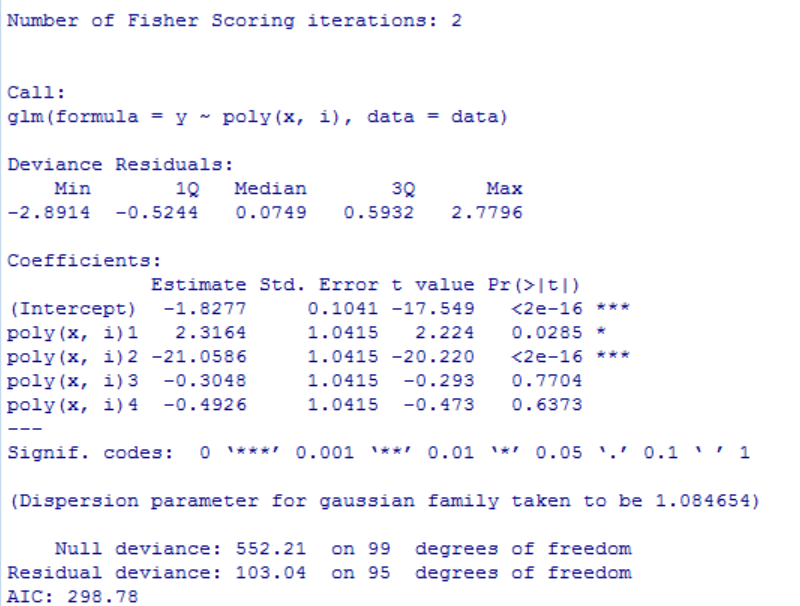
The quadratic model is the one with the smallest LOOCV error. This is to be expected from my opinion. From the equation, we know that defines Y as a second degree polynomial dependent on X.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?











*All the models agree, on the 5% confidence level.*

*The squared term is seen as significant while the reaming terms are not seem as significant at this confidence level in any model.*