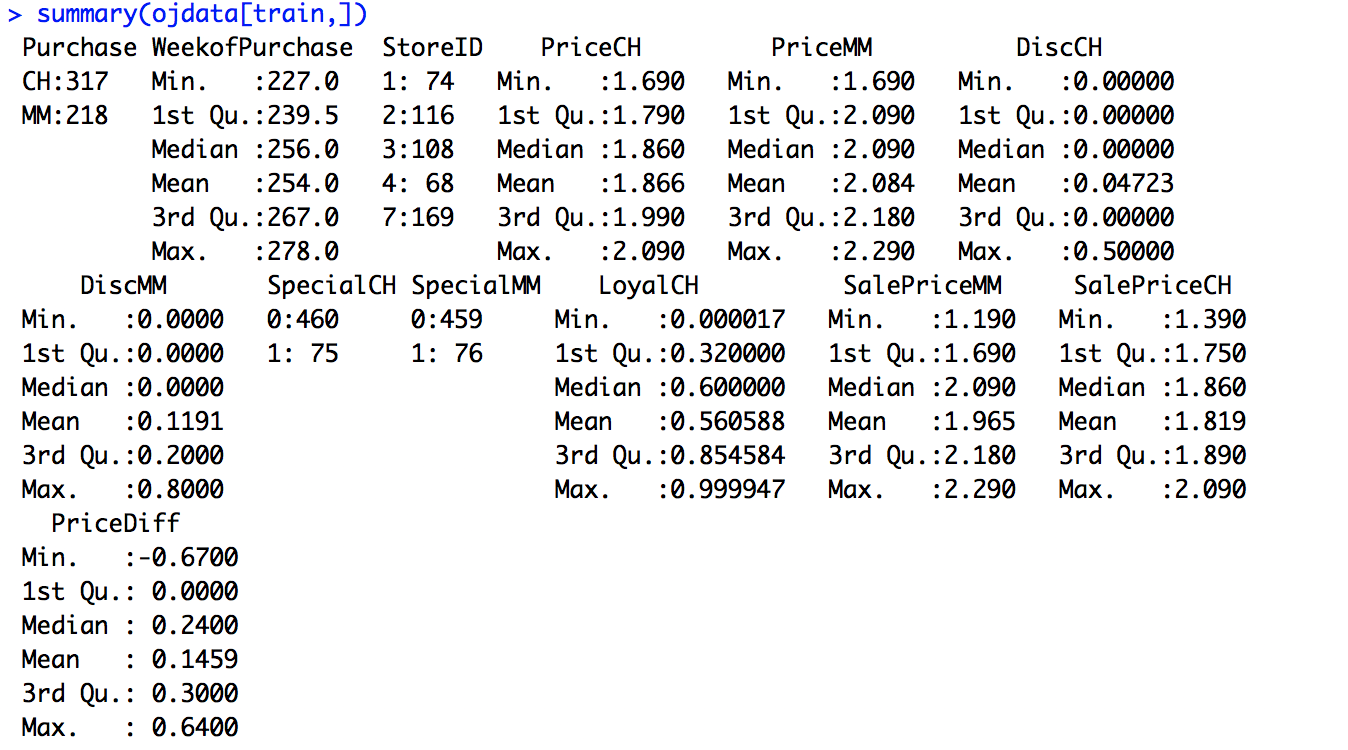
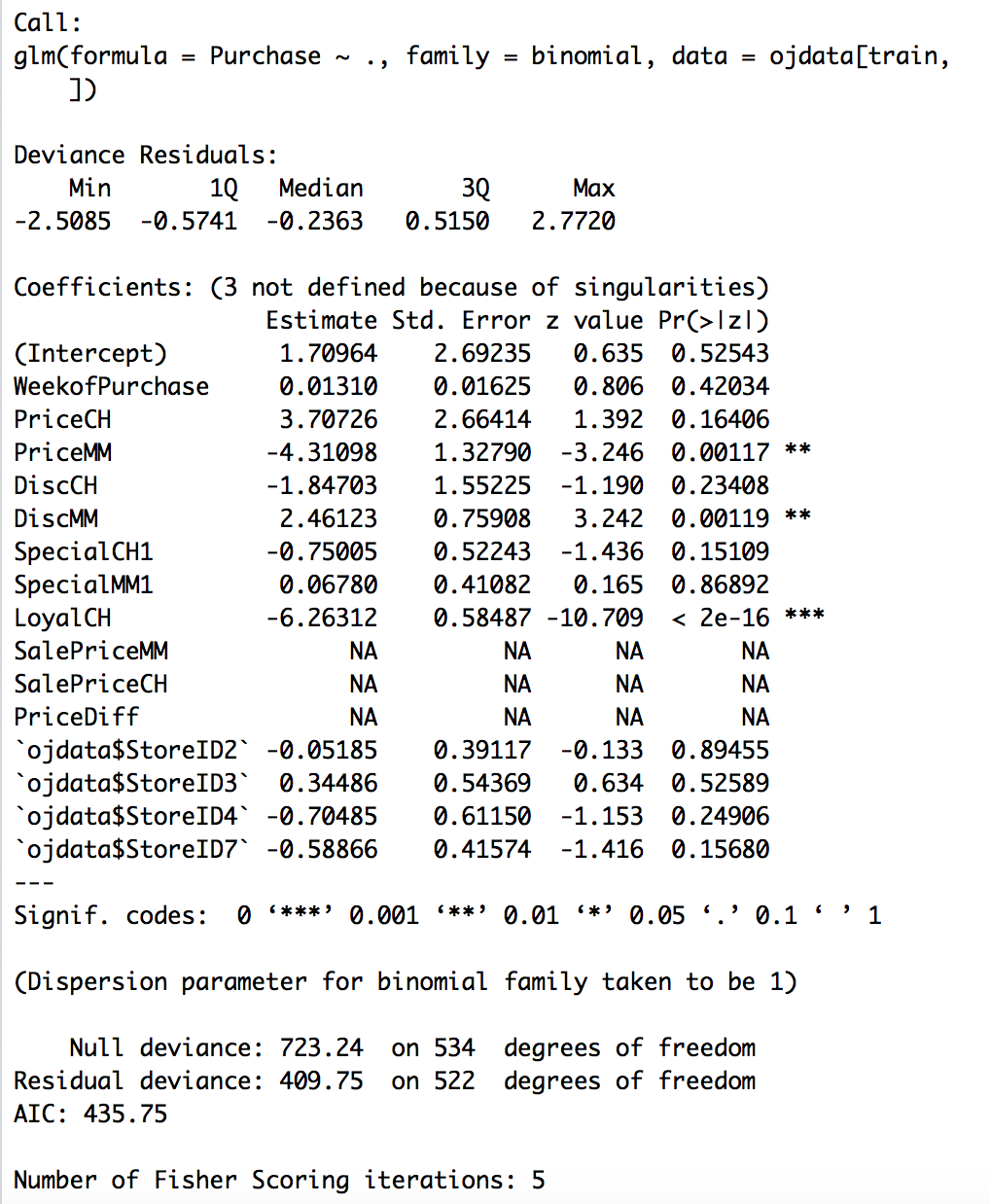
**1. Orange Juice Data**

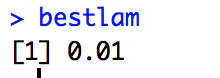
(a) Apparenttly, Purchase and StoreID variables are qualitative and need to be transformed to factor.

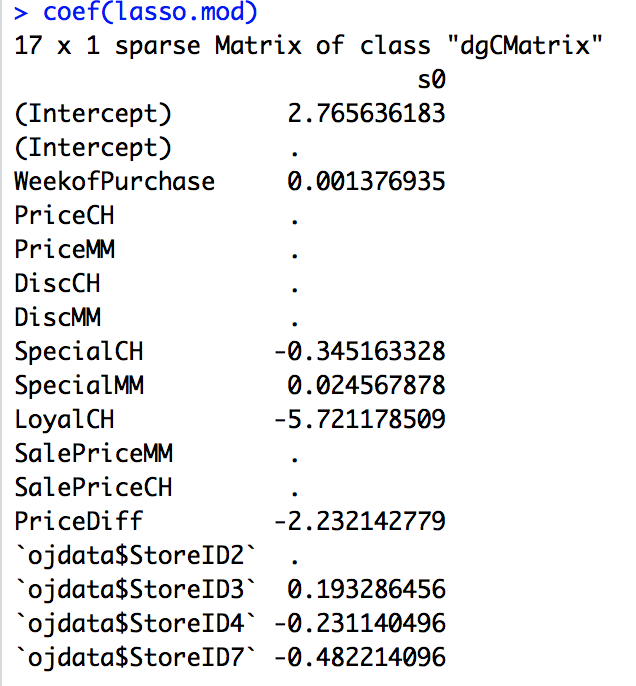


(b) The intercept is 1.7096, and coefficient for covariates are listed below. We can see that coefficient of WeekofPurchase, SpecialMM and StoreID2 are relatively close to 0, which means they have weak explanation to the target variable Purchase; while coefficient of PrichCH, PriceMM, DiscCH, DiscMM and LoyalCH are relatively far away from 0, which means they have strong explanation power.

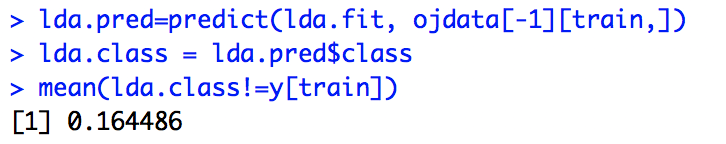
And SalePriceMM, SalePriceCH and PriceDiff has na coefficients because of singularities

(c) The best lambda is 0.01. The coefficients of the final model are listed as below.

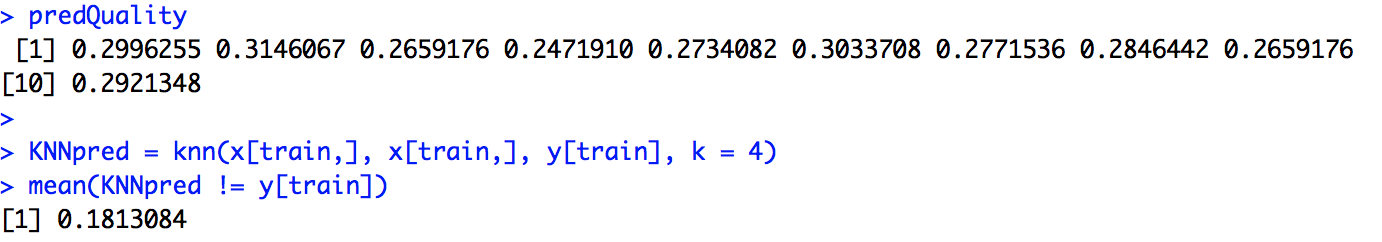




(d) The classification error on the training data is about 0.1645.



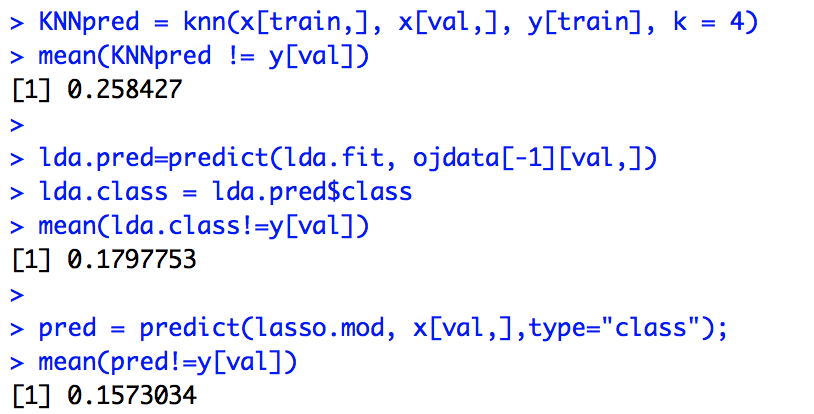
(e) K=4 is the best, with the lowest classification error validation dataset.

While the classification error on training data is 0.1813

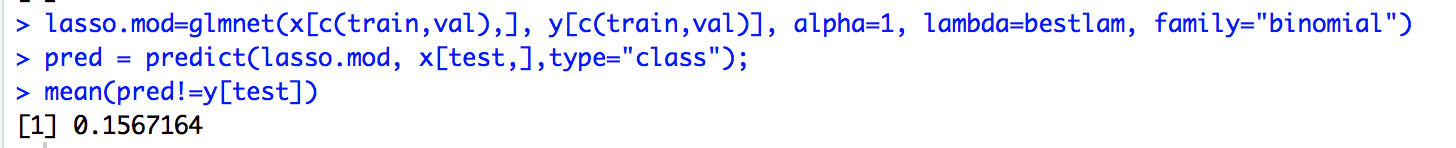
(f) I will choose Logistic Regression because of the lowest classification error on validation set. The classification error on validation set are listed as below:

KNN: 0.2584

LDA: 0.1798

Logistic Regression (Lasso): 0.1573

(g) The final classification error on test data is 0.1567



(h) People who are predicted to buy CH should receive coupon. The best threshold is 0.01 according to the graph (because the profit from true positive beats the loss from false positive), the best payoff in the test data is 577.

