Assignment4

Cameron Todd (5611230)

7 April 2017

Question 1

library("dplyr",warn.conflicts=FALSE, quietly=TRUE)  
wine.df <- read.csv("winequality.csv", header = TRUE, sep = ";")  
  
#Fit Logit model: Comment on summary  
wine\_fit1 = glm(good.quality ~ ., data = wine.df, family = binomial)  
summary(wine\_fit1)

##   
## Call:  
## glm(formula = good.quality ~ ., family = binomial, data = wine.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9878 -0.4351 -0.2207 -0.1222 2.9869   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.428e+02 1.081e+02 2.247 0.024660 \*   
## fixed.acidity 2.750e-01 1.253e-01 2.195 0.028183 \*   
## volatile.acidity -2.581e+00 7.843e-01 -3.291 0.000999 \*\*\*  
## citric.acid 5.678e-01 8.385e-01 0.677 0.498313   
## residual.sugar 2.395e-01 7.373e-02 3.248 0.001163 \*\*   
## chlorides -8.816e+00 3.365e+00 -2.620 0.008788 \*\*   
## free.sulfur.dioxide 1.082e-02 1.223e-02 0.884 0.376469   
## total.sulfur.dioxide -1.653e-02 4.894e-03 -3.378 0.000731 \*\*\*  
## density -2.578e+02 1.104e+02 -2.335 0.019536 \*   
## pH 2.242e-01 9.984e-01 0.225 0.822327   
## sulphates 3.750e+00 5.416e-01 6.924 4.39e-12 \*\*\*  
## alcohol 7.533e-01 1.316e-01 5.724 1.04e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1269.92 on 1598 degrees of freedom  
## Residual deviance: 870.86 on 1587 degrees of freedom  
## AIC: 894.86  
##   
## Number of Fisher Scoring iterations: 6

Comment: We can clearly see that most of the variables are significant to predicting whether the wine will be of good quality, however variables citric acid, free sulfur dioxide and pH level appear to not be significant in this model. You can see about half and half of the variables have a positive or negative effect on the quality of good wine such as alcohol, sulphates, residual sugar and the fixed acidity variables all have a positive effect on the quality of the wine, it appears alcohol level has the greatest positive effect on quality of wine. The density, total sulphar dioxide, chlorides and volatile acidity have negative affects on the quality of wine. Analysis of residuals - We have a null deviance of 1270 and 1598 degrees of freedom and by adding all the terms into a logit model we can reduce the residual deviance to 870 with 1587 degrees of freedom. The AIC is at 895 but cannot determine if this is good until we start taking away variable or adding higher polynomial terms or interactions. However because this is ungrouped data we can't actually refer to the residual deviance and degrees of freedom too much and will only be looking at AIC and other predictive power metrics (later).

Q2 (A)

library("mgcv")

## Loading required package: nlme

##   
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':  
##   
## collapse

## This is mgcv 1.8-17. For overview type 'help("mgcv-package")'.

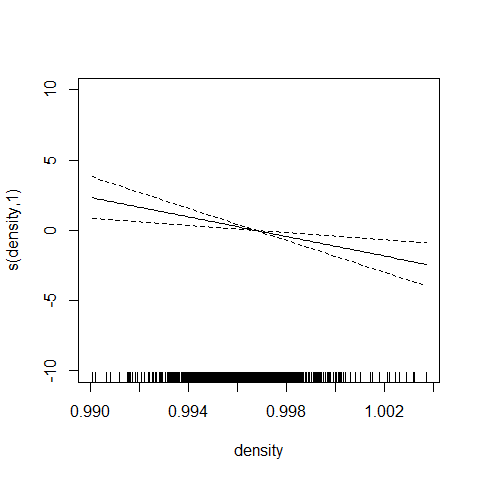
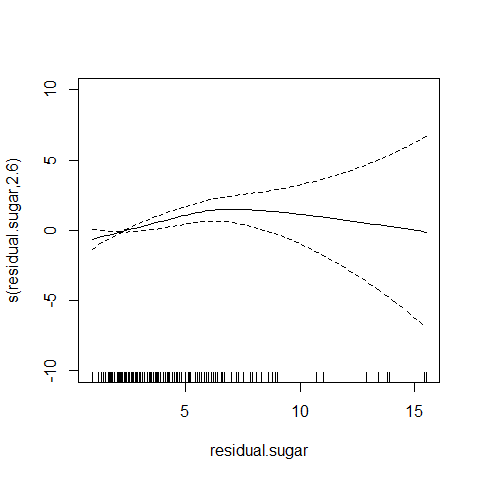
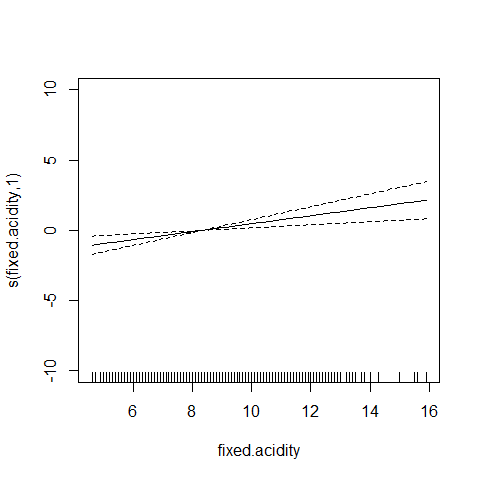
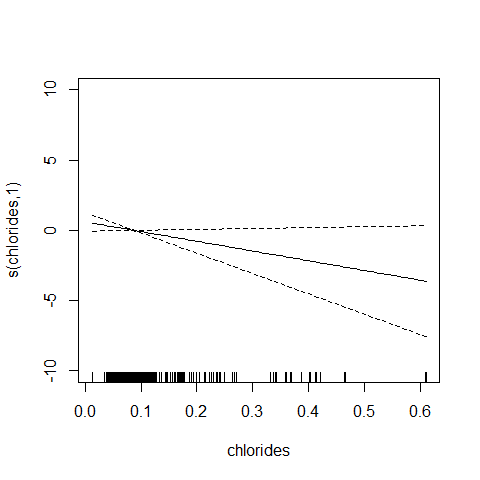
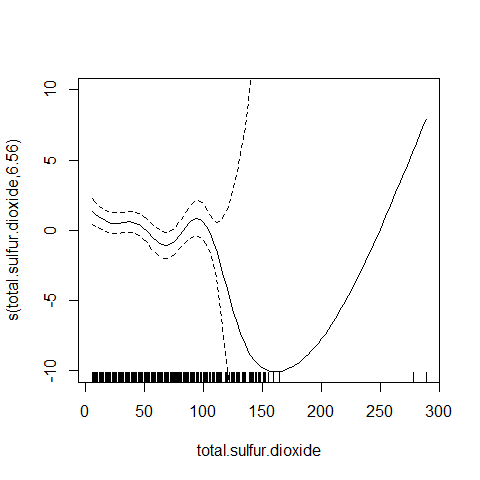
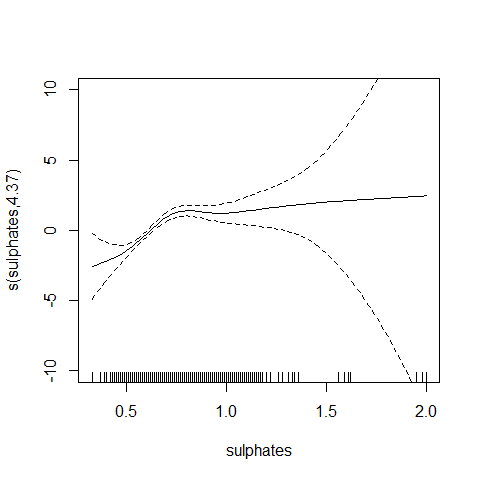
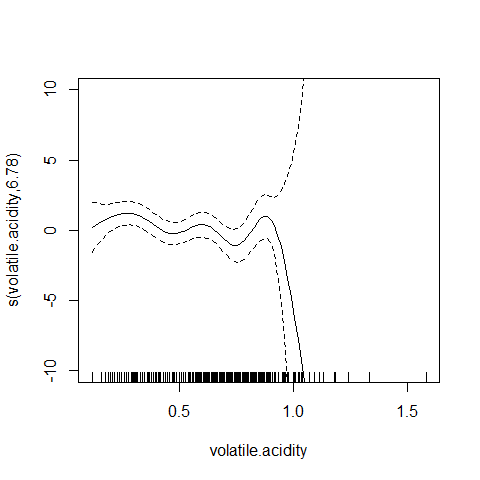
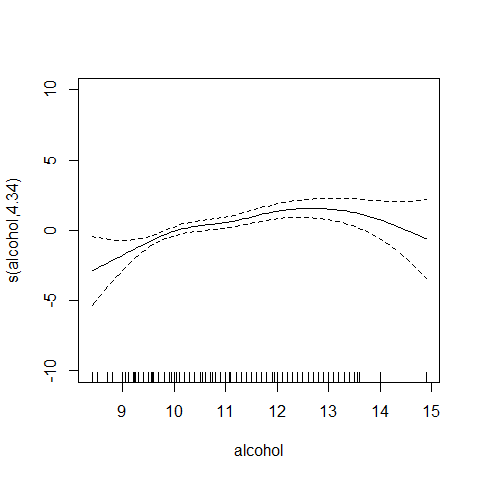
full\_model = glm(good.quality ~ ., data = wine.df, family = binomial)  
null\_model = glm(good.quality ~ 1, data = wine.df, family = binomial)  
  
#Stepwise Regression  
step(null\_model, scope = formula(full\_model), direction = "both", trace = FALSE)

##   
## Call: glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
## density, family = binomial, data = wine.df)  
##   
## Coefficients:  
## (Intercept) alcohol volatile.acidity   
## 226.7504 0.7823 -2.9128   
## sulphates total.sulfur.dioxide chlorides   
## 3.6987 -0.0136 -8.4408   
## fixed.acidity residual.sugar density   
## 0.2812 0.2328 -240.9465   
##   
## Degrees of Freedom: 1598 Total (i.e. Null); 1590 Residual  
## Null Deviance: 1270   
## Residual Deviance: 872.1 AIC: 890.1

basic\_model = glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
 total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
 density, family = binomial, data = wine.df)  
summary(basic\_model)

##   
## Call:  
## glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
## density, family = binomial, data = wine.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0158 -0.4314 -0.2220 -0.1255 2.9883   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.268e+02 9.163e+01 2.475 0.013336 \*   
## alcohol 7.823e-01 1.120e-01 6.983 2.88e-12 \*\*\*  
## volatile.acidity -2.913e+00 6.467e-01 -4.504 6.66e-06 \*\*\*  
## sulphates 3.699e+00 5.287e-01 6.997 2.62e-12 \*\*\*  
## total.sulfur.dioxide -1.360e-02 3.447e-03 -3.946 7.95e-05 \*\*\*  
## chlorides -8.441e+00 3.259e+00 -2.590 0.009593 \*\*   
## fixed.acidity 2.812e-01 8.029e-02 3.502 0.000462 \*\*\*  
## residual.sugar 2.328e-01 7.009e-02 3.322 0.000893 \*\*\*  
## density -2.409e+02 9.202e+01 -2.618 0.008835 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1269.92 on 1598 degrees of freedom  
## Residual deviance: 872.08 on 1590 degrees of freedom  
## AIC: 890.08  
##   
## Number of Fisher Scoring iterations: 6

#Fitting smoothers  
wine.gam = gam(good.quality ~ s(alcohol) + s(volatile.acidity) + s(sulphates) + s(total.sulfur.dioxide) + s(chlorides) + s(fixed.acidity) + s(residual.sugar) + s(density), data = wine.df, family = binomial)  
plot(wine.gam, ylim = c(-10,10))



#Fit model with polynomials  
poly\_model = glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
 total.sulfur.dioxide + chlorides + fixed.acidity + poly(residual.sugar,2.6) +   
 density, family = binomial, data = wine.df)  
summary(poly\_model)

##   
## Call:  
## glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + poly(residual.sugar,   
## 2.6) + density, family = binomial, data = wine.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0133 -0.4371 -0.2201 -0.1288 3.0085   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.854e+02 9.675e+01 2.950 0.00318 \*\*   
## alcohol 6.970e-01 1.198e-01 5.819 5.93e-09 \*\*\*  
## volatile.acidity -3.039e+00 6.541e-01 -4.646 3.38e-06 \*\*\*  
## sulphates 3.728e+00 5.290e-01 7.047 1.83e-12 \*\*\*  
## total.sulfur.dioxide -1.351e-02 3.405e-03 -3.967 7.27e-05 \*\*\*  
## chlorides -8.944e+00 3.362e+00 -2.661 0.00780 \*\*   
## fixed.acidity 2.946e-01 8.045e-02 3.663 0.00025 \*\*\*  
## poly(residual.sugar, 2.6)1 1.275e+01 5.079e+00 2.511 0.01205 \*   
## poly(residual.sugar, 2.6)2 -9.179e+00 5.713e+00 -1.607 0.10813   
## density -2.983e+02 9.693e+01 -3.078 0.00209 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1269.92 on 1598 degrees of freedom  
## Residual deviance: 868.08 on 1589 degrees of freedom  
## AIC: 888.08  
##   
## Number of Fisher Scoring iterations: 6

#New model with interaction 2way  
interaction2model = glm(formula = good.quality ~ (alcohol+volatile.acidity+sulphates+total.sulfur.dioxide+chlorides+fixed.acidity+residual.sugar+density)^2, family = binomial, data = wine.df)  
  
step(null\_model, scope = formula(interaction2model), direction = "both", trace= FALSE)

##   
## Call: glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
## density + total.sulfur.dioxide:chlorides + alcohol:sulphates +   
## total.sulfur.dioxide:fixed.acidity + alcohol:fixed.acidity +   
## alcohol:density + fixed.acidity:density, family = binomial,   
## data = wine.df)  
##   
## Coefficients:  
## (Intercept) alcohol   
## 2.407e+03 -2.252e+02   
## volatile.acidity sulphates   
## -2.485e+00 -1.130e+01   
## total.sulfur.dioxide chlorides   
## -8.368e-03 1.349e+01   
## fixed.acidity residual.sugar   
## 3.946e+01 2.452e-01   
## density total.sulfur.dioxide:chlorides   
## -2.448e+03 -8.684e-01   
## alcohol:sulphates total.sulfur.dioxide:fixed.acidity   
## 1.464e+00 6.690e-03   
## alcohol:fixed.acidity alcohol:density   
## -2.939e-01 2.285e+02   
## fixed.acidity:density   
## -3.624e+01   
##   
## Degrees of Freedom: 1598 Total (i.e. Null); 1584 Residual  
## Null Deviance: 1270   
## Residual Deviance: 816.2 AIC: 846.2

interaction2stepw = glm(good.quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +   
 chlorides + fixed.acidity + residual.sugar + density + total.sulfur.dioxide:chlorides +   
 alcohol:sulphates + total.sulfur.dioxide:fixed.acidity +   
 alcohol:fixed.acidity + alcohol:density + fixed.acidity:density, family = binomial, data = wine.df)  
summary(interaction2stepw)

##   
## Call:  
## glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
## density + total.sulfur.dioxide:chlorides + alcohol:sulphates +   
## total.sulfur.dioxide:fixed.acidity + alcohol:fixed.acidity +   
## alcohol:density + fixed.acidity:density, family = binomial,   
## data = wine.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.44220 -0.40250 -0.18566 -0.07111 2.99169   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 2.407e+03 7.733e+02 3.113 0.001852  
## alcohol -2.252e+02 6.006e+01 -3.750 0.000177  
## volatile.acidity -2.485e+00 6.767e-01 -3.672 0.000241  
## sulphates -1.130e+01 6.402e+00 -1.765 0.077499  
## total.sulfur.dioxide -8.368e-03 1.751e-02 -0.478 0.632677  
## chlorides 1.349e+01 4.945e+00 2.729 0.006357  
## fixed.acidity 3.946e+01 2.276e+01 1.734 0.082902  
## residual.sugar 2.452e-01 8.330e-02 2.943 0.003248  
## density -2.448e+03 7.792e+02 -3.142 0.001679  
## total.sulfur.dioxide:chlorides -8.684e-01 1.935e-01 -4.488 7.21e-06  
## alcohol:sulphates 1.464e+00 5.876e-01 2.492 0.012709  
## total.sulfur.dioxide:fixed.acidity 6.690e-03 2.394e-03 2.794 0.005199  
## alcohol:fixed.acidity -2.939e-01 6.066e-02 -4.846 1.26e-06  
## alcohol:density 2.285e+02 6.064e+01 3.767 0.000165  
## fixed.acidity:density -3.624e+01 2.266e+01 -1.599 0.109710  
##   
## (Intercept) \*\*   
## alcohol \*\*\*  
## volatile.acidity \*\*\*  
## sulphates .   
## total.sulfur.dioxide   
## chlorides \*\*   
## fixed.acidity .   
## residual.sugar \*\*   
## density \*\*   
## total.sulfur.dioxide:chlorides \*\*\*  
## alcohol:sulphates \*   
## total.sulfur.dioxide:fixed.acidity \*\*   
## alcohol:fixed.acidity \*\*\*  
## alcohol:density \*\*\*  
## fixed.acidity:density   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1269.92 on 1598 degrees of freedom  
## Residual deviance: 816.18 on 1584 degrees of freedom  
## AIC: 846.18  
##   
## Number of Fisher Scoring iterations: 7

#New model with 3 way interaction - However this model results in the same AIC as a 2way interaction and therefore i won't choose this model  
interaction3model = glm(formula = good.quality ~ (alcohol+volatile.acidity+sulphates+total.sulfur.dioxide+chlorides+fixed.acidity+residual.sugar+density)^3, family = binomial, data = wine.df)  
  
step(null\_model, scope = formula(interaction3model), direction = "both", trace = FALSE)

##   
## Call: glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
## density + total.sulfur.dioxide:chlorides + alcohol:sulphates +   
## total.sulfur.dioxide:fixed.acidity + alcohol:fixed.acidity +   
## alcohol:density + fixed.acidity:density, family = binomial,   
## data = wine.df)  
##   
## Coefficients:  
## (Intercept) alcohol   
## 2.407e+03 -2.252e+02   
## volatile.acidity sulphates   
## -2.485e+00 -1.130e+01   
## total.sulfur.dioxide chlorides   
## -8.368e-03 1.349e+01   
## fixed.acidity residual.sugar   
## 3.946e+01 2.452e-01   
## density total.sulfur.dioxide:chlorides   
## -2.448e+03 -8.684e-01   
## alcohol:sulphates total.sulfur.dioxide:fixed.acidity   
## 1.464e+00 6.690e-03   
## alcohol:fixed.acidity alcohol:density   
## -2.939e-01 2.285e+02   
## fixed.acidity:density   
## -3.624e+01   
##   
## Degrees of Freedom: 1598 Total (i.e. Null); 1584 Residual  
## Null Deviance: 1270   
## Residual Deviance: 816.2 AIC: 846.2

interaction3stepw = glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
 total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
 density + total.sulfur.dioxide:chlorides + alcohol:sulphates +   
 total.sulfur.dioxide:fixed.acidity + alcohol:fixed.acidity +   
 alcohol:density + fixed.acidity:density, family = binomial,   
 data = wine.df)  
summary(interaction3stepw)

##   
## Call:  
## glm(formula = good.quality ~ alcohol + volatile.acidity + sulphates +   
## total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar +   
## density + total.sulfur.dioxide:chlorides + alcohol:sulphates +   
## total.sulfur.dioxide:fixed.acidity + alcohol:fixed.acidity +   
## alcohol:density + fixed.acidity:density, family = binomial,   
## data = wine.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.44220 -0.40250 -0.18566 -0.07111 2.99169   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 2.407e+03 7.733e+02 3.113 0.001852  
## alcohol -2.252e+02 6.006e+01 -3.750 0.000177  
## volatile.acidity -2.485e+00 6.767e-01 -3.672 0.000241  
## sulphates -1.130e+01 6.402e+00 -1.765 0.077499  
## total.sulfur.dioxide -8.368e-03 1.751e-02 -0.478 0.632677  
## chlorides 1.349e+01 4.945e+00 2.729 0.006357  
## fixed.acidity 3.946e+01 2.276e+01 1.734 0.082902  
## residual.sugar 2.452e-01 8.330e-02 2.943 0.003248  
## density -2.448e+03 7.792e+02 -3.142 0.001679  
## total.sulfur.dioxide:chlorides -8.684e-01 1.935e-01 -4.488 7.21e-06  
## alcohol:sulphates 1.464e+00 5.876e-01 2.492 0.012709  
## total.sulfur.dioxide:fixed.acidity 6.690e-03 2.394e-03 2.794 0.005199  
## alcohol:fixed.acidity -2.939e-01 6.066e-02 -4.846 1.26e-06  
## alcohol:density 2.285e+02 6.064e+01 3.767 0.000165  
## fixed.acidity:density -3.624e+01 2.266e+01 -1.599 0.109710  
##   
## (Intercept) \*\*   
## alcohol \*\*\*  
## volatile.acidity \*\*\*  
## sulphates .   
## total.sulfur.dioxide   
## chlorides \*\*   
## fixed.acidity .   
## residual.sugar \*\*   
## density \*\*   
## total.sulfur.dioxide:chlorides \*\*\*  
## alcohol:sulphates \*   
## total.sulfur.dioxide:fixed.acidity \*\*   
## alcohol:fixed.acidity \*\*\*  
## alcohol:density \*\*\*  
## fixed.acidity:density   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1269.92 on 1598 degrees of freedom  
## Residual deviance: 816.18 on 1584 degrees of freedom  
## AIC: 846.18  
##   
## Number of Fisher Scoring iterations: 7

library("R330")

## Loading required package: s20x

## Loading required package: leaps

## Loading required package: rgl

## Loading required package: lattice

#Analyse the predictive power of the model with 3 tests below  
#Cross Validation and Bootstrap results  
cross.val(interaction2stepw)

## Mean Specificity = 0.9647697   
## Mean Sensitivity = 0.3599562   
## Mean Correctly classified = 0.8824843

err.boot(interaction2stepw)

## $err  
## [1] 0.1131957  
##   
## $Err  
## [1] 0.1133171

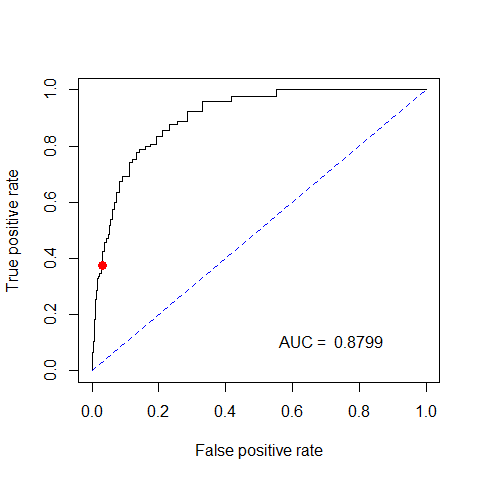
cross.val(poly\_model)

## Mean Specificity = 0.963975   
## Mean Sensitivity = 0.3453792   
## Mean Correctly classified = 0.8795597

err.boot(poly\_model)

## $err  
## [1] 0.1175735  
##   
## $Err  
## [1] 0.1177123

#ROC curve  
ROC.curve(good.quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +   
 chlorides + fixed.acidity + residual.sugar + density + total.sulfur.dioxide:chlorides +   
 alcohol:sulphates + total.sulfur.dioxide:fixed.acidity +   
 alcohol:fixed.acidity + alcohol:density + fixed.acidity:density, family = binomial, data = wine.df)

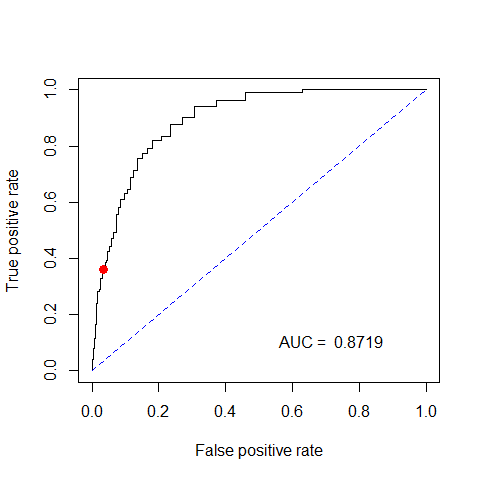


## Area under ROC curve = 0.8799

R330::HLstat(interaction2stepw)

## Value of HL statistic = 7.661   
## P-value = 0.467

ROC.curve(good.quality ~ alcohol + volatile.acidity + sulphates +   
 total.sulfur.dioxide + chlorides + fixed.acidity + poly(residual.sugar,2.6) +   
 density, family = binomial, data = wine.df)

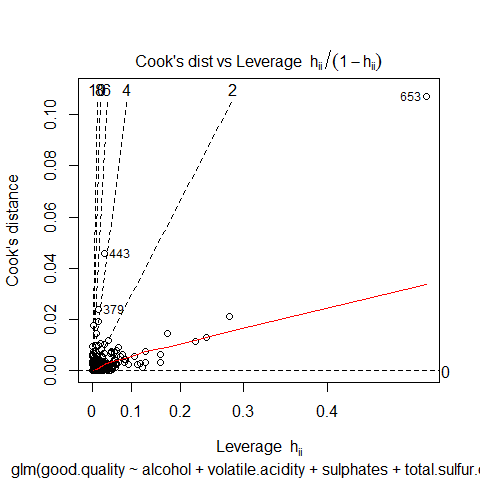
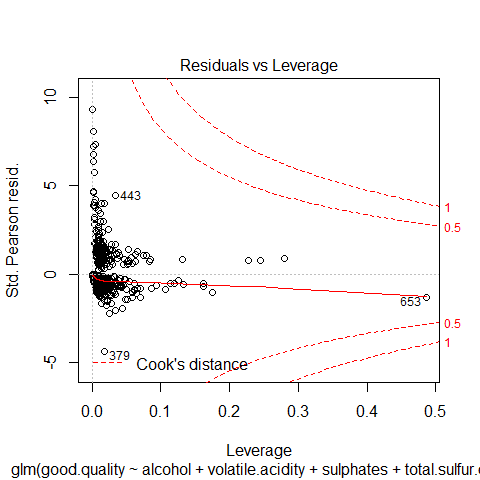
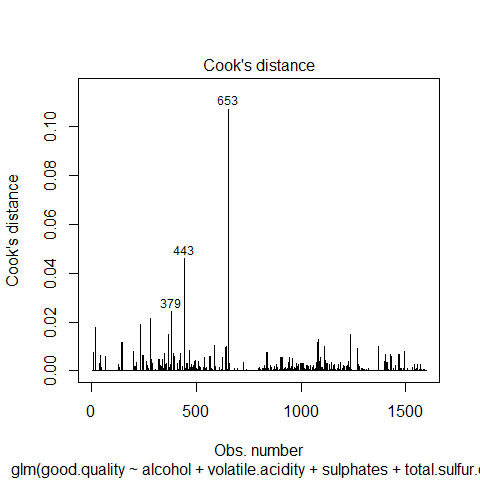
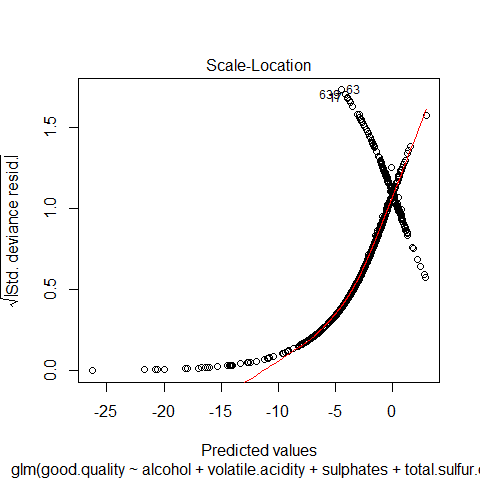
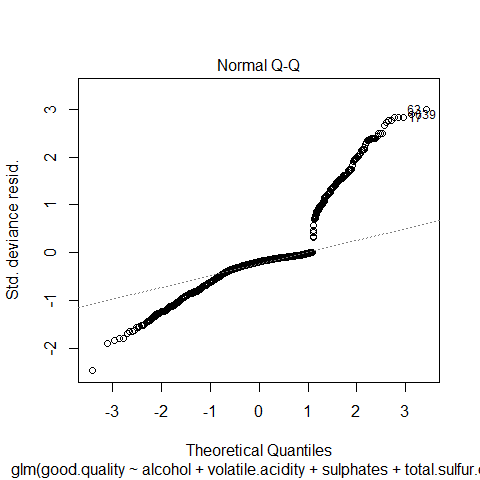
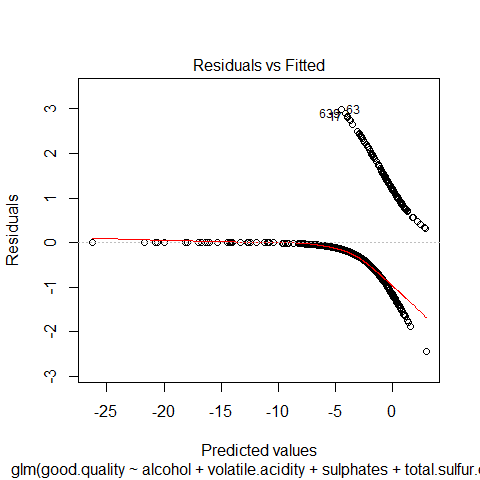


## Area under ROC curve = 0.8719

R330::HLstat(poly\_model)

## Value of HL statistic = 6.938   
## P-value = 0.543

#Diagnostics  
plot(interaction2stepw, which = 1:6)



Q2 (A) In total i fitted 4 different model. the first was the standard basic linear model with all variables and then took out variables that were not significant (basic\_model). I then tested for any polynomials that could be added onto the basic model by fitting smoothers over each variable, a few variable suggested high polynomial terms (4 and above), this would of over smoothed my model and would not be good for new data so i only added a polynomial term of 2 to the variable residual.sugar. After comapring these two models i could see the polynomial helped with a slightly lower AIC from 890 to 888. I then created 2 new models with interaction terms upto 2 and the 2nd model was upto 3, i then used the stepwise function of bothways to choose the optimum amount of terms in both models by naturally minimising AIC. The 2way and 3way interaction models pretty much came out the same and the AIC was inseperable so i choose the 2way interaction model.

I then completed a analysis over the predictive power over the 2 models i am looking at closest the polynomial model and the 2-way interaction model. As you can see above, both have a p-value greater then 0.05 in the Hosmer-Lemeshow statistic, both models also have a high CV and bootstrap results and ROC curve. However the 2way interaction model slightly fits better and has a lower AIC, therefore i will go with this model. I will go into more detai on predictive power in 2B.

I then completed the relevant diagnostics (above) I was a bit concerned with observation 653 but I had no real evidence that the observation was wrong and that by deleting it we would increase the regression coefficients or make fitted probabilities more extreme and we would end up overstating the predictive ability.

All in all i think the 2-way interaction model is the best one here out of the other 5.

pred = predict(interaction2stepw, type = "response")  
predcode = ifelse(pred < 0.5, 0, 1)  
table(with(wine.df,good.quality),predcode)

## predcode  
## 0 1  
## FALSE 1337 45  
## TRUE 136 81

specificity = 1337/(1337 + 45)  
Sensitivity = 81/(136 + 81)  
print(specificity)

## [1] 0.9674385

print(Sensitivity)

## [1] 0.3732719

Q2(B) When determining how well the model is able to predict (goodness of fit) whether the wine is of good quality or not, i will analyse the ROC curve, the confusion matrix, cross validation/bootstrap results and the Hosmer-Lemeshow statistic. Hosmer-Lemeshow statistic result as printed above of 0.467 is above the 0.05 which normally indicated problems in the fittted logistic model. Cross validation/bootstrap results show that our Bootstrap mean correctly classified = 1 - 0.113 = 0.887, 88.7% of observations, this is the average error result accross the 10 tests, you can also see the specificity and sensitivity results which are similar to the ROC explained below. The ROC curve amd confusion matrix above returns similar results showing what's the split of type's of observations we are more likely to predict well or not i.e. Sensitivity and Specificity with at this stage a defaul c value of 0.5 meaning if the observation has a probability of greater then 0.5 then we would predict it to be a good bottle of wine. The ROC curve and confusion matrix shows a sensitivity value of 0.37 and specificty of 1 - 0.033 = .967. The specificty result is great this means our model predicts with an accuracy of 96.7% that a wine bottle is of bad quality when the case is truly a bad wine, however our sensitivity is only .37 meaning with a probability of .37 of predicting a good wine when the actual case is a good wine, these are with the c-value of 0.5 (default). The area under the curve of 0.8799 is also means we can predict out of all observations with an accuracy of 88%, this agrees with CV/bootstrap. Overall the model seems pretty good but can obviously change the weighting of c if we care about predicting good bottles of wine more accurately.

Q3 (A)

pred = predict(interaction2stepw, type = "response")  
pred.success = pred[wine.df$good.quality == 1]  
quantile(pred.success,0.1)

## 10%   
## 0.06978707

Q3 (B) substitute the 0.069 into the matrix instead of 0.5. Makes the model probabilities closer to 1 meaning we have a higher chance of predicting good bottles of wine.

pred = predict(interaction2stepw, type = "response")  
predcode = ifelse(pred < 0.069788, 0, 1)  
table(with(wine.df,good.quality),predcode)

## predcode  
## 0 1  
## FALSE 965 417  
## TRUE 22 195

specificity = 963/(963 + 419)  
Sensitivity = 195/(195 + 22)  
Error\_rate = (417 + 22)/(965+417+22+195)  
print(specificity)

## [1] 0.6968162

print(Sensitivity)

## [1] 0.8986175

print(Error\_rate)

## [1] 0.2745466

After finding the optimum c-value to predict within 90% accuracy of a good bottle of wine, the c-value being 0.0697. The new Sensitivity value is 0.898 or rounding up to 90% as requested but because of the trade-off effect the specificty has dropped from the 96% to 69%. This means that our model will now predict with a probability of 90% of predicting a good wine when the case is truely a good wine and with a probability of 69% predict a bad wine when the true case is a bad wine. The overall in-sample error is 0.274 or 27.4%. Obviously comparing these figures to the default of 0.5 means that now any observations with a predicted probability of greater then 0.069 is now considered potentially a good wine instead of any probability over 0.5 we will say is potentially a good wine so of course logically our specificty was going to increase by .32 and obviously the trade-off effect reduces the sensitivity by .07 to 89.8%. Because the change in c was around trying to save any good wine being given away, the probability of giving away a good wine has significantly reduced to 10% instead of the 1-.37 = 63% chance liklihood with the default of c so year i agree this change in c is a good change.

Q3 (C)

winetest.df <- read.csv("winequality-test.csv", header = TRUE, sep = ";")  
predict(interaction2stepw, newdata = winetest.df, type = "response")

## 1 2 3 4 5 6   
## 0.011596620 0.022392350 0.378737150 0.001596446 0.899022430 0.004901514   
## 7 8 9 10   
## 0.239320092 0.001658700 0.275967358 0.159307693

Obviously with our newly constructed c value of 0.06978 we will happily give away any bottles with a probability of less then this figure and keep any bottles above this figure, therefore; bottles we will give away are observations 1,2,4,6, and 8 Bottles we will keep are observations 3,5,7,9,10