Homework 1 – Generalised Linear Models

1. Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient. Build a single predictor logistic regression model (fit.single) using X_h as the predictor. Write the equations relating the dependent variable (Response) to the explanatory variable in terms of:

(Highest Estimated predictor : CategoryPhotograph)

Category predictor equations :

a. Probabilities: P (Y=Yes | X_h=x)

Probability(Y=Yes,X_h = x) = 1/ 1 + $e^{-(0.1929+14.3731*Xh)}$

where (X_h = CategoryPhotography)

b. Odds: P (Y=Yes)

Probability(Y=Yes)/1- Probability(Y=Yes) = $e^{(0.1929+14.3731*Xh)}$

where (Xh = CategoryPhotography)

c. Logit

Log(Probability(Y=Yes)/1- Probability(Y=Yes)) = 0.1929 + 14.3731 * Xh

where (Xh = CategoryPhotography)

2. Write the estimated equation for the fit.all model in all three formats (if the number of predictors is more than four, then include only those four predictors whose absolute value estimates are the highest):

a. The logit as a function of the predictors.

Equation for logit = 0.45 + 13.4192 * CategoryPhotography + 1.54 * CategoryEverythingElse + 1.07 * CurrencyGDP + 1.05* EndDayMon

b. The odds as a function of the predictors.

Equation for odds is =

$e^{(0.45+13.4129*CategoryPhotography+1.54*CategoryEverythingElse +1.07*CurrencyGDP+1.12*EndDayMon)}$

c. The probability as a function of the predictors

Probability as a function of predictors = Equation for odds / 1 + Equation for odds

$(e^{(0.45+13.4129*CategoryPhotography+1.54*CategoryEverythingElse +1.07*CurrencyGDP+1.12*EndDayMon)})$ /

$(1 + e^{(0.45+13.4129*CategoryPhotography+1.54*CategoryEverythingElse +1.07*CurrencyGDP+1.12*EndDayMon)})$

= 1 /

$1 + e^{-(0.45+13.4129*CategoryPhotography+1.54*CategoryEverythingElse +1.07*CurrencyGDP+1.12*EndDayMon)}$

3. Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient in the fit.all. Compute the odds ratio that estimated a single unit increase in X_h, holding the other predictors constant. For example, if X_h=1 then: odds(X_1+1,X_2,…,Xq)/ odds(X_1,X_2,…,Xq) =

Provide the interpretation for this regression coefficient. If it were a linear regression model, how would the interpretation change for a single unit increase in X_h.

From question 1 we know that the highest estimated predictor is CategoryPhotograph

odds(X_1+1,X_2,…,Xq)/ odds(X_1,X_2,…,Xq) =

$e^{\text{sum for all other predictor terms+coefficient of CategoryPhotograph*(categoryPhotograph + 1)}}$

(we consider X_h = CategoryPhotograph because it has highest estimated predictor)

odds(X_1+1,X_2,…,Xq)/ odds(X_1,X_2,…,Xq) =

$e^{\text{ coefficient of CategoryPhotograph}}$

(we consider X_h = CategoryPhotograph because it has highest estimated predictor)

Thus we notice coefficient of CategoryPhotograph = 13.4192

odds(X_1+1,X_2,…,Xq)/ odds(X_1,X_2,…,Xq) = $e^{13.4725}$ = 672797.720

For the change in linear regression change for a single unit increase will eventually make a change of predictor coefficient in the prediction of the model. i.e the overall prediction. So our value of 14.58 would change.

For logistic regression it would only lead to change in predictor coefficient in log odds value.

4. Build a reduced logistic regression model (fit.reduced) using only the predictors that are statistically significant. Assess if the reduced model is equivalent to the full model. Justify your answer.

The anova test is used to compare the two given models in the regression. We get a very low p-value 0.000641 for the test and hence can reject the null hypothesis which stats the models are equivalent and hence we choose the alternate hypothesis and confirm that the two models are not equivalent

5. Compute the dispersion of your model and run the dispersion diagnostic test. If the constructed model is overdispersed, then discuss the ways to deal with the issue.

We calculate the residual deviance as = 1203.7 – (A)

WE calculate the degrees of freedom for deviance of residuals as = 1168 - (B)

The model has a dispersion = (A)/(B) = 1.03

When the test for over dispersion is run we get values for variance which is observed divided by the variance which is theoretical as = 0.4494

P-value that is obtained = 1

We notice that 0.4494 is not very different when compared statistically to 1

We can conclude that the model is not over-dispersed.

If overdispersion was present we need to refit using quasi-binomial distribution