Homework6

Anshuman Narayan

11/28/2018

Q1

Q1a

To conduct a likelihood ratio to test if $\beta_1 = \beta_2 = 0$, we need to build the null model corresponding to this test and a full model that would prove that this null hypothesis is false. We state the null hypothesis and alternate hypothesis as follows: $H_0 : \beta_1 = \beta_2 = 0$ $H_a : H_0$ is false

```
full_mod=glm(formula = atleastone ~ width + weight,data=hc,family=binomial)
null_mod<-update(full_mod,~.-width)
null_mod<-update(null_mod,~.-weight)
devfull=deviance(full_mod)
devnull=deviance(null_mod)
g_sq=devnull-devfull
cat("\nThe g-squared variable is ",g_sq)
```

```
##
## The g-squared variable is  32.86664
```

```
cat("\nThe p-value of this variable following a Chi-squared distribution",1-pchisq(g_sq,1))
```

```
##
## The p-value of this variable following a Chi-squared distribution 9.870277e-09
```

From the value of p we can see it is less than the 5 % significance level we set for our test. Thus, we can reject the null hypothesis that weight and width cannot be removed from the model. ####Q1b To conduct a likelihood ratio to test if $\beta_1 = 0$, we build a null model that reflects this hypothesis and an alternative hypothesis that is a full model. We state the null and alternate hypothesis as follows: $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

```
null_mod<-update(full_mod,~.-width)
devnull=deviance(null_mod)
g_sq=devnull-devfull
cat("\nThe g-squared variable is ",g_sq)
```

```
##
## The g-squared variable is  2.845263
```

```
cat("\nThe p-value of this variable following a Chi-squared distribution",1-pchisq(g_sq,1))
```

```
##
## The p-value of this variable following a Chi-squared distribution 0.09164361
```

From the value of p above we can see that it is greater than the 5% significance level we set for our test. Thus, we fail to reject the null hypothesis that width does not have an effect on the prediction of the atleastone variable.

Q1c

To conduct a likelihood ratio to test if $\beta_2 = 0$, we need to build a null model corresponding to this test and a full model that would prove that this null hypothesis is false. We state the null hypothesis and the alternative hypothesis as follows: $H_0 : \beta_2 = 0$ $H_a : \beta_2 \neq 0$

```
null_mod<-update(full_mod,~.-weight)
devnull=deviance(null_mod)
g_sq=devnull-devfull
cat("\nThe g-squared variable is ",g_sq)
```

```
##
## The g-squared variable is  1.560777
```

```
cat("\nThe p-value of this variable following a Chi-squared distribution",1-pchisq(g_sq,1))
```

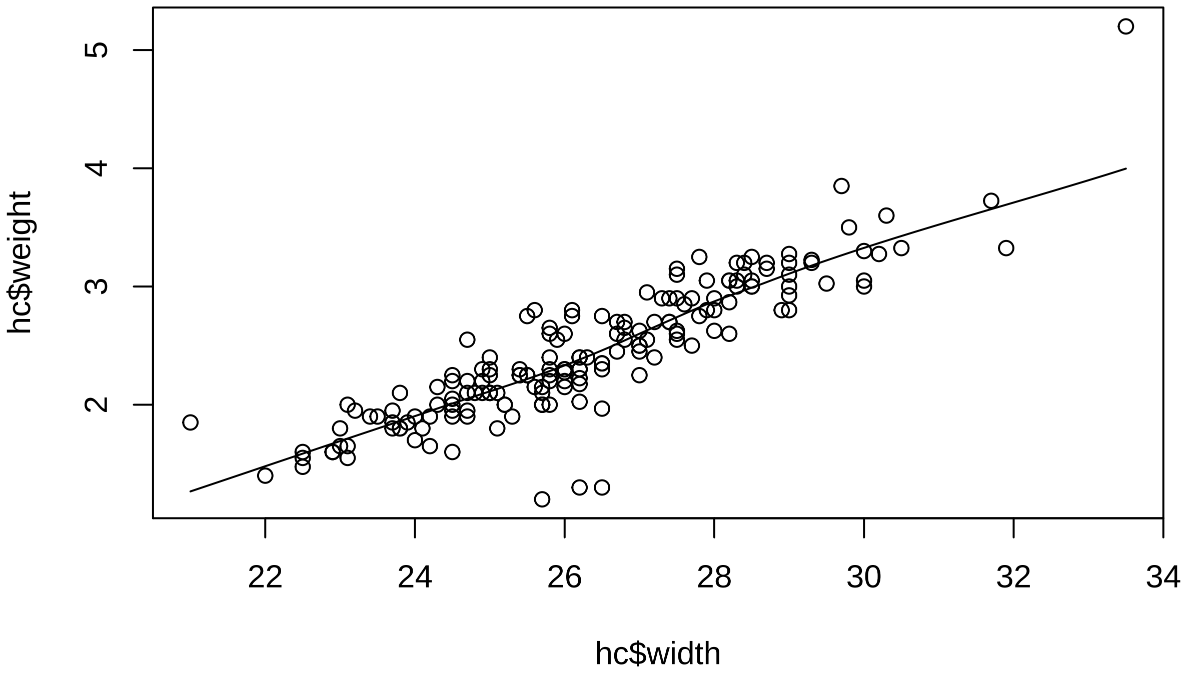
```
##
## The p-value of this variable following a Chi-squared distribution 0.2115515
```

From the value of p above we can see that it is greater than the 5% significance level we set for our test. Thus, we fail to reject the null hypothesis that weight does not have an effect on the prediction of the atleastone variable.

Q1d

The p-value of 1(a) shows strong evidence against the null hypothesis while tests 1(b) and 1(c) do not. We look at the scatter plot of weight vs width.

```
scatter.smooth(x=hc$width,y=hc$weight)
```



We can see that weight and width have a positive correlation. This means that they have a similar effect on the predicitive model. Thus removing either one fails to reject the null hypothesis. But on removing both the effect is removed which is significant to the model.

Q2

We build the model as specified in the question

```
mod=glm(formula=atleastone~ width + weight + color + weight*color,data=hc,family=binomial)
summary(full_mod)
```

```
##
## Call:
## glm(formula = atleastone ~ width + weight, family = binomial,
##     data = hc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1127  -1.0344   0.5304   0.9006   1.7207
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.3547     3.5280  -2.652  0.00801 **
## width         0.3068     0.1819   1.686  0.09177 .
## weight        0.8338     0.6716   1.241  0.21445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 192.89  on 170  degrees of freedom
## AIC: 198.89
##
## Number of Fisher Scoring iterations: 4
```

Q2a

We start the backward elimination approach using the drop1 function on the full model and updating the model and repeating it till all variables are significant to the model.

```
drop1(mod,test="Chi")
```

```
## Single term deletions
##
## Model:
## atleastone ~ width + weight + color + weight * color
##              Df Deviance    AIC    LRT Pr(>Chi)
## <none>            179.33 197.33
## width         1   181.66 197.66 2.3251  0.12730
## weight:color  3   186.21 198.21 6.8800  0.07582 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("The 'width' variable has p-value larger than the 10% significant level.")
```

```
## The 'width' variable has p-value larger than the 10% significant level.
```

```
modn<-update(mod,~.-width)
drop1(modn,test="Chi")
```

```
## Single term deletions
##
## Model:
```

```
## atleastone ~ weight + color + weight:color
##               Df Deviance    AIC   LRT Pr(>Chi)
## <none>             181.66 197.66
## weight:color  3    188.54 198.54 6.886  0.07562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("No variables have p-values larger than the 10% significant level.")
```

```
## No variables have p-values larger than the 10% significant level.
```

```
summary(modn)
```

```
##
## Call:
## glm(formula = atleastone ~ weight + color + weight:color, family = binomial,
##     data = hc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0875  -0.8766   0.5412   0.8399   1.9421
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.1868     2.2999  -0.516   0.6058
## weight                    0.1947     1.0303   0.189   0.8501
## colordarkMed             -6.7299     3.4353  -1.959   0.0501 .
## colorlightMed            -0.4335     5.4046  -0.080   0.9361
## colormedium              -1.2654     2.5847  -0.490   0.6244
## weight:colordarkMed       3.5601     1.5634   2.277   0.0228 *
## weight:colorlightMed      0.8536     2.1551   0.396   0.6920
## weight:colormedium        1.2149     1.1419   1.064   0.2874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 181.66  on 165  degrees of freedom
## AIC: 197.66
##
## Number of Fisher Scoring iterations: 5
```

Q2b

```
pred=predict(modn,newdata=data.frame(weight=2.35,color='medium'),se.fit=TRUE)
z_perc=qnorm(0.95)
LB=pred$fit-z_perc*pred$se.fit
UB=pred$fit+z_perc*pred$se.fit
ilogit=function(u) return (exp(u)/(1+exp(u)));
ilogit(cbind(LB,UB))
```

```
##          LB        UB
## 1 0.6133194 0.7789459
```

Q2c

```
confint.default(modn,level=0.95)
```

```
##                             2.5 %      97.5 %
## (Intercept)            -5.6945933  3.320971128
## weight                 -1.8246629  2.214139206
## colordarkMed          -13.4629482  0.003156193
## colorlightMed         -11.0263876 10.159416276
## colormedium            -6.3313982  3.800526324
## weight:colordarkMed     0.4958824  6.624258590
## weight:colorlightMed   -3.3703783  5.077579692
## weight:colormedium     -1.0232032  3.453010033
```

Q2d

The selected model is given by the following formula $log(\pi(x)/(1-\pi(x))) = \beta_0+\beta_1 weight+\beta_2 colordarkMed+\beta_3 colorlightMed+\beta_4 colormedium+\beta_5 weight : colordarkMed+\beta_6 weight : colorlightMed+\beta_7 weight : colormedium$ where β_0 to β_7 are given in the coefficients displayed above from intercept to weight:colormedium respectively.

When color is darkMed then every unit increase in the weight changes the log odds of the crab having at least satellite by $\beta_1+\beta_2+\beta_5 = (0.1947381-6.7298960+3.5600705) = -2.9750874$ or multiplies the odds of crab having at least one satellite by $exp(\beta_1+\beta_2+\beta_5) = exp(-2.9750874)$

When color is lightmedium, then every unit increase in the weight changes the log odds of the crab having at least one satellite by: $\beta_1+\beta_3+\beta_6 = (0.1947381-0.4334857+0.8536007) = 0.6148531$ or multiplies the odds of crab having at least one satellite by $\beta_1+\beta_3+\beta_6 = exp(0.6148531)$

When color is medium, then every unit increase in the weight changes the log odds of the crab having at least one satellite by: $\beta_1+\beta_4+\beta_7 = (0.1947381-1.2654359+1.2149034 = 0.1442056$ or multiplies the odds of crab having at least one satellite by $\beta_1+\beta_4+\beta_7 = exp(0.1442056)$

Q3

```
aids<-read.table('aids.txt')
```

Q3a

The subjects version of this model assumes that $y_1, ..., y_4$, the observed sample proportions of successes

```
aids$symptomsYes/(aids$symptomsYes+aids$symptomsNo)
```

```
## [1] 0.1308411 0.2831858 0.1746032 0.2181818
```

are a realization of $Y_1, ..., Y_4$ which are independent and $n_i Y_i$~$Binom[n,π(x_i)], i=1,…,4 $ where, using variable names notation, $log(\pi(x_i)/(1-\pi_i)) = \beta_0+\beta_1 * race+\beta_2 * aztUse+\beta_3 race : aztUse$ and $n_1, ..., n_4$ are

```
aids$symptomsYes+aids$symptomsNo
```

```
## [1] 107 113  63  55
```

We fit this model using

```
mod=glm(formula=cbind(symptomsYes,symptomsNo)~race*aztUse,data=aids,family=binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = cbind(symptomsYes, symptomsNo) ~ race * aztUse,
##     family = binomial, data = aids)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.2763     0.3265  -3.909 9.26e-05 ***
## racew              0.3476     0.3875   0.897    0.370
## aztUseyes         -0.2771     0.4655  -0.595    0.552
## racew:aztUseyes   -0.6878     0.5852  -1.175    0.240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.3499e+00  on 3  degrees of freedom
## Residual deviance: 1.4655e-14  on 0  degrees of freedom
## AIC: 25.476
##
## Number of Fisher Scoring iterations: 3
```

Q2b

```
step(mod,direction="backward")
```

```
## Start:  AIC=25.48
## cbind(symptomsYes, symptomsNo) ~ race * aztUse
##
##                Df Deviance    AIC
## - race:aztUse  1   1.3835 24.860
## <none>             0.0000 25.476
##
## Step:  AIC=24.86
## cbind(symptomsYes, symptomsNo) ~ race + aztUse
##
##           Df Deviance    AIC
## - race    1   1.4206 22.897
## <none>        1.3835 24.860
## - aztUse  1   8.2544 29.731
##
## Step:  AIC=22.9
## cbind(symptomsYes, symptomsNo) ~ aztUse
##
##           Df Deviance    AIC
## <none>        1.4206 22.897
## - aztUse  1   8.3499 27.826
```

```
##
## Call:  glm(formula = cbind(symptomsYes, symptomsNo) ~ aztUse, family = binomial,
##     data = aids)
##
```

```
## Coefficients:
## (Intercept)    aztUseyes
##     -1.0361      -0.7218
##
## Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
## Null Deviance:       8.35
## Residual Deviance: 1.421     AIC: 22.9
```

Q2c

Since this data is grouped, we can use the deviance to assess goodness of fit. We set hypothesis that H_0: $\beta_1 = 0$ H_a: null model is false.

```
reduced_mod<-update(mod,~.-race-race:aztUse)
summary(reduced_mod)
```

```
##
## Call:
## glm(formula = cbind(symptomsYes, symptomsNo) ~ aztUse, family = binomial,
##     data = aids)
##
## Deviance Residuals:
##       1        2        3        4
## -0.4813   0.5102   0.6026  -0.7521
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0361     0.1755  -5.904 3.54e-09 ***
## aztUseyes    -0.7218     0.2787  -2.590  0.00961 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.3499  on 3  degrees of freedom
## Residual deviance: 1.4206  on 2  degrees of freedom
## AIC: 22.897
##
## Number of Fisher Scoring iterations: 4
```

```
full_mod<-deviance(mod)
reduced_mod<-deviance(reduced_mod)
g_sq=reduced_mod-full_mod
g_sq
```

```
## [1] 1.420614
```

```
cat("The probability that g-squared follows a Chi-squared distribution is ",1-pchisq(g_sq,1))
```

```
## The probability that g-squared follows a Chi-squared distribution is  0.2333024
```

From the p-value of g-squared we can say that the we fail to reject the null hypothesis.

Q3d

We have selected the model with the estimates $log\hat{\pi}(x)/(1-\hat{\pi}(x)) = \beta_0+\beta_1*aztUse = -1.0361-0.7218*aztUse$

The estimated odds of symptomsYes for aztUseYes is exp(-1.0361-0.7218) = 0.1724 The estimated odds of symptomsYes for aztUseNo is exp(-1.0361) = 0.3548358 The estimated odds ratio between aztUse(yes,no) and symptoms(yes,no) is exp(-0.7218) = 0.4858769 = 0.1724/0.3548358 The estimated odds ratio between aztUse(no,yes) and symptoms(yes,no) is exp(0.7218) = 2.058135 = 0.3548358/0.1724