STA 6443 - HW4 solution

Exercise 1.

(a) After performing stepwise selection below, we can conclude that the best set of predictors for a logistic regression model predicting whether a female is a liver patient includes two numerical variables: DB and Aspartate.

```
liverF = liver[which(liver$Gender == "Female"),]

glm.null.F <- glm(LiverPatient ~ 1, data = liverF, family = "binomial")
glm.full.F <- glm(LiverPatient ~ Age+TB+DB+Alkphos+Alamine+Aspartate+TP+ALB,
data = liverF, family = "binomial")

# Perform stepwise selection based on AIC criteria
glm.liverF<-step(glm.null.F, scope = list(upper=glm.full.F),
                      direction="both",test="Chisq", trace = F)
```

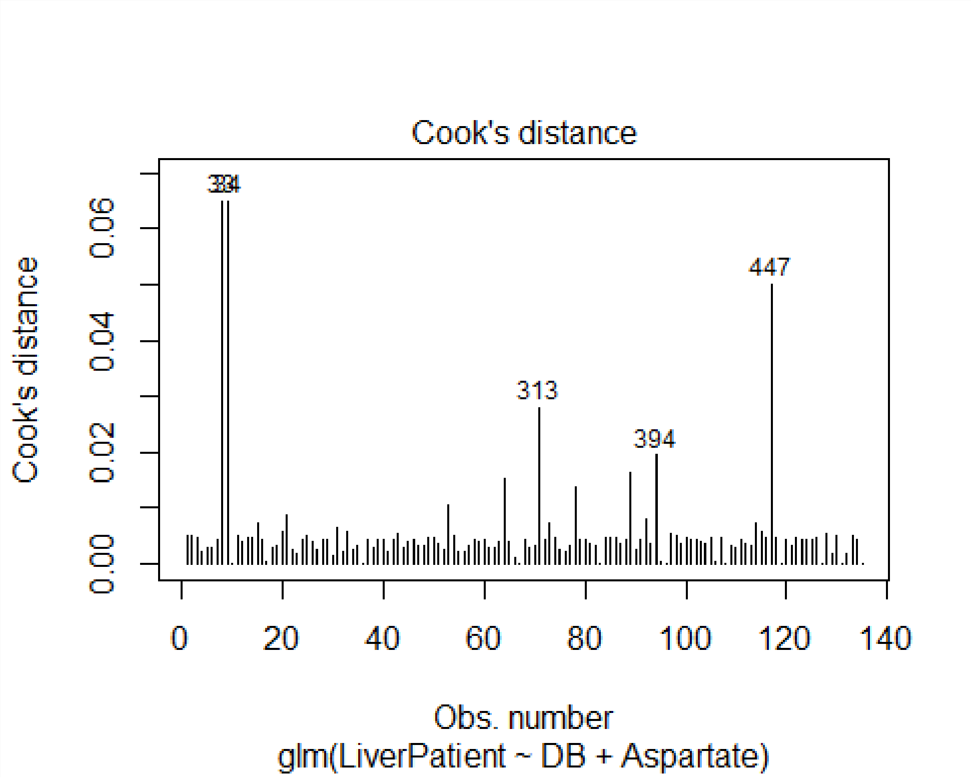
(b) In the model, DB and Aspartate are both significant with p-values less than the significance level 0.1. The result of HL (Hosmer-Lemeshow) test below has a p-value of 0.45 which is greater than 0.1. Thus we accept the null hypothesis and conclude that the model fit the data well.

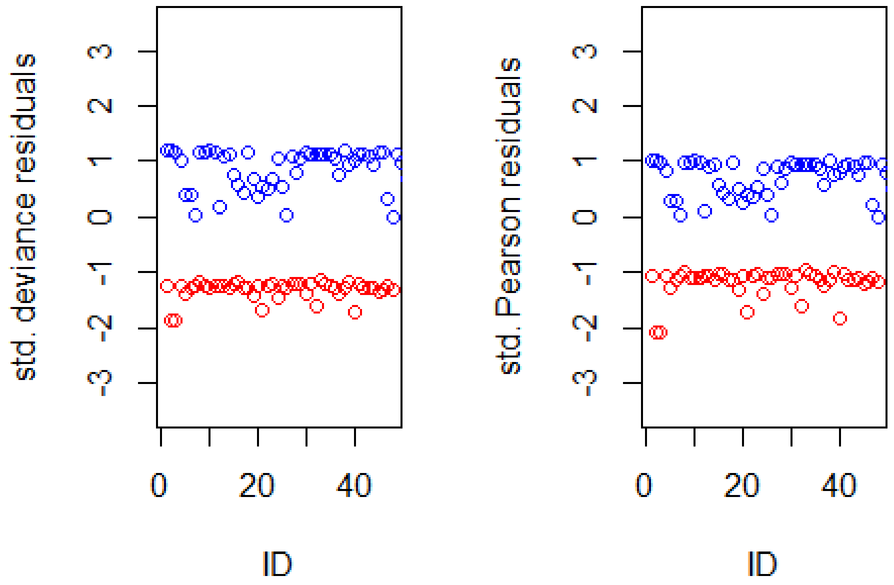
As presented in the following Influence Diagnostics plot, the highest cook's d is around 0.07 which is less than the threshold 0.25. Thus, there is no unduly influential point. So, there is no need for refitting the model. The residual plots do not show any systematic pattern and there is not observations with very large residuals. Thus our model assumption seems valid.

```
summary(glm.liverF)

##
## Call:
## glm(formula = LiverPatient ~ DB + Aspartate, family = "binomial",
##     data = liverF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8178  -1.2223   0.4402   1.1091   1.2049
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.32480    0.31013  -1.047   0.2950
## DB           0.94479    0.55808   1.693   0.0905 .
## Aspartate    0.01106    0.00616   1.796   0.0726 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 175.72  on 134  degrees of freedom
## Residual deviance: 154.27  on 132  degrees of freedom
## AIC: 160.27
##
## Number of Fisher Scoring iterations: 7

# cook's d
plot(glm.liverF, which = 4, id.n = 5)
```

```
hoslem.test(glm.liverF$y, fitted(glm.liverF), g=10)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  glm.liverF$y, fitted(glm.liverF)
## X-squared = 7.7535, df = 8, p-value = 0.4579
```





(c) The estimated Odds Ratio (OR) for DB and Aspartate are 2.57 and 1.01, respectively. This means that, for each unit increasing of DB, there will be 2.57 (=exp(0.94)) times increasing of odds, and for each unit increasing of Aspartate, there will be 1.011 (=exp(0.011)) times increasing of odds of an adult female being a liver patient.

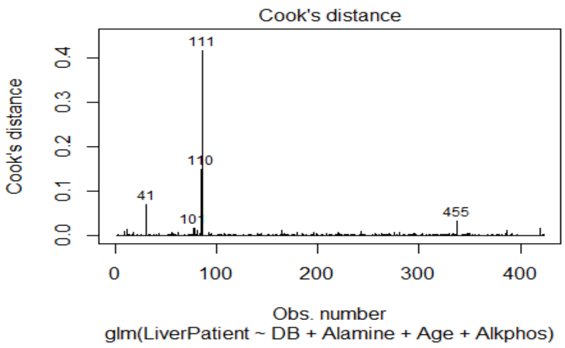
```
OR=exp(glm.liverF$coefficients)
round(OR,3)

## (Intercept)          DB   Aspartate
##       0.723       2.572       1.011
```

Exercise 2.

(a) According to the summary output of stepwise selection below, we can conclude that the best set of predictors for a logistic regression model predicting whether a male is a liver patient are: DB, Alamine, Age and Alkphos

```
liverM = liver[which(liver$Gender == "Male"),]

glm.null.M <- glm(LiverPatient ~ 1, data = liverM, family = "binomial")
glm.full.M <- glm(LiverPatient ~ Age+TB+DB+Alkphos+Alamine+Aspartate+TP+ALB,
data = liverM, family = "binomial")

# Perform stepwise selection based on AIC criteria
glm.liverM <- step(glm.null.M, scope = list(upper=glm.full.M),direction="both
",test="Chisq", trace = F)
```

```
summary(glm.liverM)

##
## Call:
## glm(formula = LiverPatient ~ DB + Alamine + Age + Alkphos, family = "binom
ial",
##     data = liverM)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3405  -0.5170   0.3978   0.8614   1.3756
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.476570   0.481336  -3.068  0.00216 **
## DB           0.512503   0.176066   2.911  0.00360 **
## Alamine      0.016218   0.005239   3.095  0.00197 **
## Age          0.020616   0.008095   2.547  0.01087 *
## Alkphos      0.001740   0.001058   1.645  0.09992 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 476.28  on 422  degrees of freedom
## Residual deviance: 395.05  on 418  degrees of freedom
## AIC: 405.05
##
## Number of Fisher Scoring iterations: 7

# cook's d
plot(glm.liverM, which = 4, id.n = 5)
```



(b) As presented in the following Influence Diagnostics plot, the highest cook's d is greater than the threshold 0.25. Thus we refit the model without the high influential point.

(Solution interprets the refitted model, but there will be no deduction of points for the interpretation of the original model with influential points. But detection of influential observation should be addressed)

All predictors (DB, Alamine, Age, and Alkphos) in the final refitted model without influential points are significant with all their p-value less than 0.1. The result of HL test below has a p-value of 0.467 which is greater than 0.1. Thus we accept the null hypothesis and conclude that the model fit the data well.

The cook's d plot after removing the high influential point, and this time, there is no high influential point showed in the plot. Residual plots does not show any problematic patterns or large standardized residuals, thus model assumption seems valid.

```
glm.liverM2 = glm(LiverPatient ~ DB+Alamine+Age+Alkphos, data = liverM[-inf.i
d, ], family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

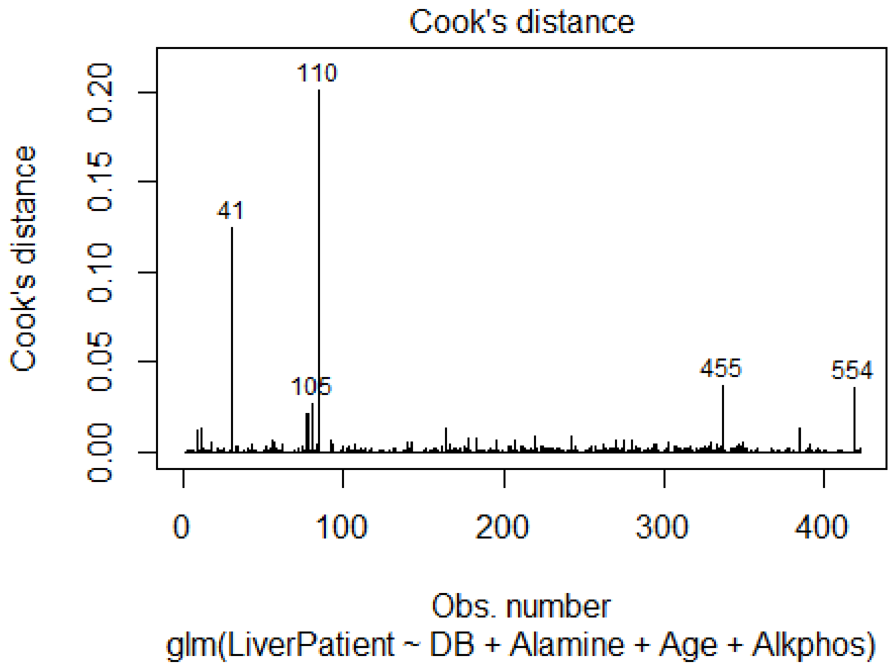
```
summary(glm.liverM2)
```

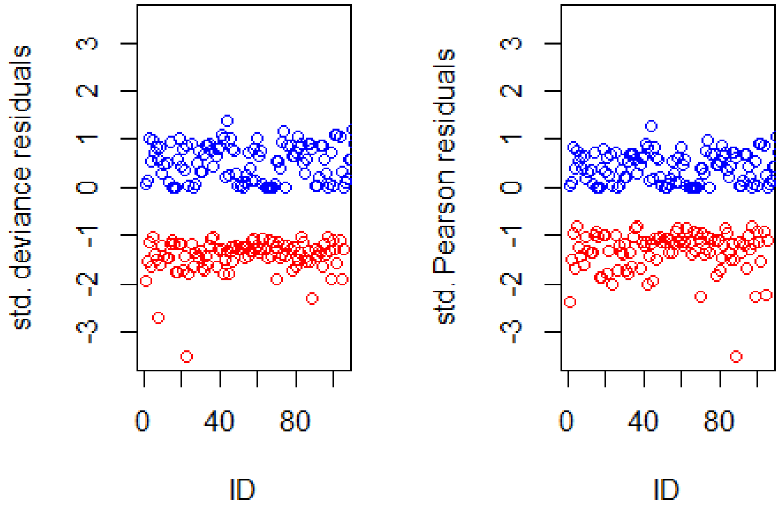
```
##
## Call:
## glm(formula = LiverPatient ~ DB + Alamine + Age + Alkphos, family = "binom
ial",
##     data = liverM[-inf.id, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5166   0.0000   0.3301   0.8648   1.4696
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.902754   0.527386  -3.608 0.000309 ***
## DB           0.573104   0.198893   2.881 0.003958 **
## Alamine      0.015850   0.005466   2.900 0.003737 **
## Age          0.020418   0.008210   2.487 0.012883 *
## Alkphos      0.003744   0.001477   2.534 0.011262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 473.51  on 421  degrees of freedom
## Residual deviance: 381.31  on 417  degrees of freedom
## AIC: 391.31
##
## Number of Fisher Scoring iterations: 8
```

```
hoslem.test(glm.liverM2$y, fitted(glm.liverM2), g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
```

```
## data:  glm.liverM2$y, fitted(glm.liverM2)
## X-squared = 7.6642, df = 8, p-value = 0.4669

# cook's d
plot(glm.liverM2, which=4, id.n=5)
```





(c) The estimation of OR for DB is 1.774. This means that, for each unit increasing of DB, there will be 1.774(=exp(0.573)) times increasing of odds of an adult male being a liver patient. The estimation of OR for Alamine is 1.016. This means that, for each unit increasing of Alamine, there will be 1.016 (=exp(0.016)) times increasing of odds of an adult male being a liver patient. The estimation of OR for Age is 1.021. This means that, for each unit increasing of Age, there will be 1.021(=exp(0.02)) times increasing of odds of an adult male being a liver patient. The estimation of OR for Alkphos is 1,004. This means that, for each unit increasing of Alkphos, there will be 1.004 (=exp(0.003)) times increasing of odds of an adult male being a liver patient.

```
OR=exp(glm.liverM2$coefficients)
round(OR,3)

## (Intercept)          DB     Alamine         Age     Alkphos
##       0.149       1.774       1.016       1.021       1.004
```

Exercise 3.

(a) The best model from stepwise selection via AIC criteria contains brainweight, totalsleep, sleepexposureindex, and predationindex.

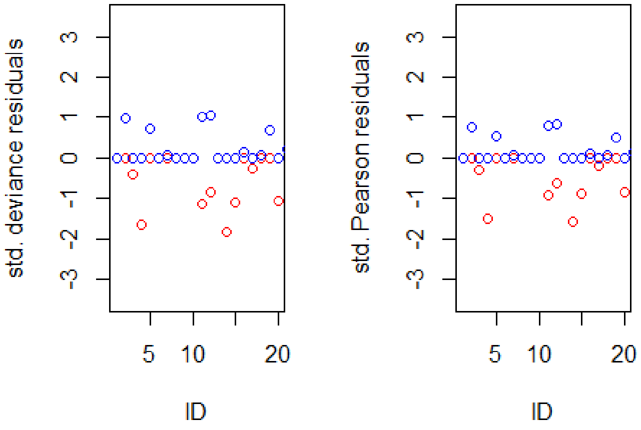
```
glm.null.sleep1 <- glm(maxlife10 ~ 1, data = sleep, family = "binomial")

glm.full.sleep1 <- glm(maxlife10 ~ bodyweight+brainweight+totalsleep+gestatio
ntime
                       +as.factor(predationindex)+as.factor(sleepexposureinde
x), data = sleep, family = "binomial")

glm.sleep1 <- step(glm.null.sleep1, scope = list(upper=glm.full.sleep1),
                   direction="both",test="Chisq", trace = F)

summary(glm.sleep1)

##
## Call:
## glm(formula = maxlife10 ~ brainweight + totalsleep + as.factor(sleepexposu
reindex) +
##     as.factor(predationindex), family = "binomial", data = sleep)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.42528  -0.00004   0.00000   0.00013   2.37523
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -6.602e+00  4.864e+00  -1.357   0.1747
## brainweight                    5.101e-02  5.084e-02   1.003   0.3157
## totalsleep                     4.230e-01  2.647e-01   1.598   0.1100
## as.factor(sleepexposureindex)2  4.998e+00  2.559e+00   1.953   0.0508 .
## as.factor(sleepexposureindex)3  3.636e+01  9.624e+03   0.004   0.9970
## as.factor(sleepexposureindex)4  3.370e+01  1.037e+04   0.003   0.9974
## as.factor(sleepexposureindex)5  7.341e+01  1.262e+04   0.006   0.9954
## as.factor(predationindex)2     -2.535e+00  1.960e+00  -1.293   0.1960
## as.factor(predationindex)3     -2.512e+01  1.253e+04  -0.002   0.9984
## as.factor(predationindex)4     -1.826e+01  6.795e+03  -0.003   0.9979
## as.factor(predationindex)5     -5.264e+01  1.143e+04  -0.005   0.9963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.31  on 50  degrees of freedom
## Residual deviance: 15.88  on 40  degrees of freedom
## AIC: 37.88
##
## Number of Fisher Scoring iterations: 20
```

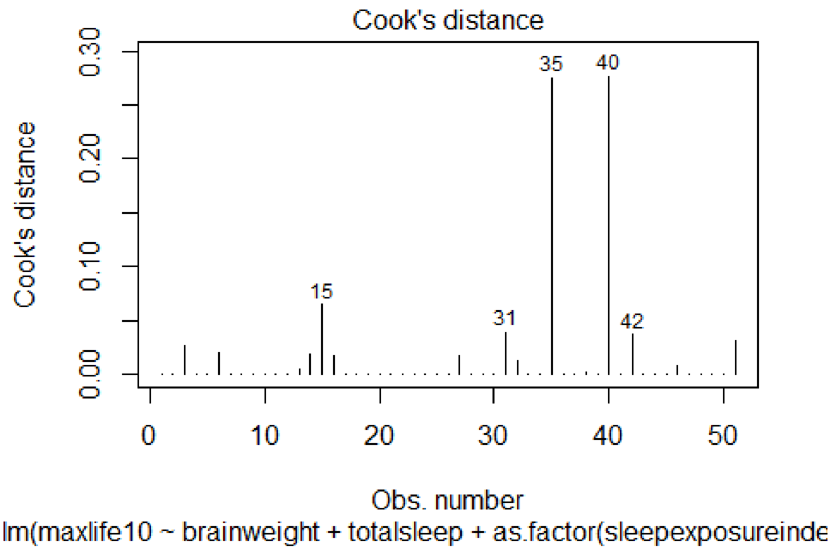
(b) Among 4 chosen predictors, only sleepexposureindex is significant with p-value for sleepexposureindex2 less than 0.1. The goodness of fit test for the model has p-value of 0.53, which indicates the model fit is adequate

In the diagnostic plots, we find two observations with relatively large cook’s d compared to others.Residual plots looks okay without problematic issues.

```
hoslem.test(glm.sleep1$y, fitted(glm.sleep1), g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  glm.sleep1$y, fitted(glm.sleep1)
## X-squared = 7.0397, df = 8, p-value = 0.5324
```





(c) We only interpret the significant one, estimated OR for sleepexposureindex2. The odds ratio is estimated as 148.05 so we can say that the odds of having maximum lifespan at least 10 years for a species with sleepexposureindex2 is 148.05 (=exp(4.99)) times of the odds for a species with sleepexposureindex1. A species under other sleepexposureindex levels (3,4, and 5) does not have significantly different odds comapred to a species with sleepexposureindex level1.

```
OR=exp(glm.sleep1$coefficients)
round(OR,3)

##                     (Intercept)                     brainweight
##                    1.000000e-03                    1.052000e+00
##                      totalsleep as.factor(sleepexposureindex)2
##                    1.527000e+00                    1.480500e+02
## as.factor(sleepexposureindex)3 as.factor(sleepexposureindex)4
##                    6.173141e+15                    4.332708e+14
## as.factor(sleepexposureindex)5     as.factor(predationindex)2
##                    7.603846e+31                    7.900000e-02
##     as.factor(predationindex)3     as.factor(predationindex)4
##                    0.000000e+00                    0.000000e+00
##     as.factor(predationindex)5
##                    0.000000e+00
```

Exercise 4.

(a) Treating the index variables as continuous, stepwise select brainweight, totalsleep, sleepexposureindex and predationindex.

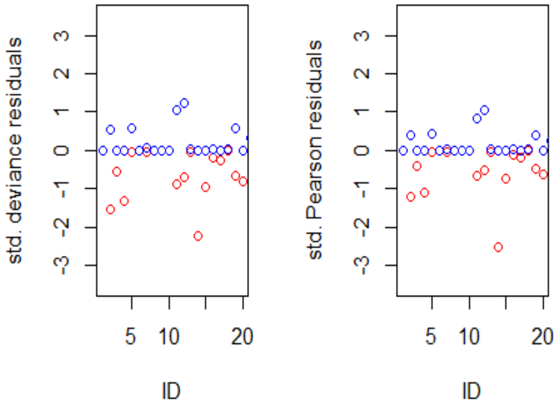
```
glm.null.sleep2 <- glm(maxlife10 ~ 1, data = sleep, family = "binomial")
glm.full.sleep2 <- glm(maxlife10 ~ bodyweight+brainweight+totalsleep+gestatio
ntime
                       + predationindex + sleepexposureindex, data = sleep, f
amily = "binomial")

glm.sleep2 <- step(glm.null.sleep2, scope = list(upper=glm.full.sleep2),
                    direction="both",test="Chisq", trace = F)

summary(glm.sleep2)

##
## Call:
## glm(formula = maxlife10 ~ brainweight + totalsleep + sleepexposureindex +
##     predationindex, family = "binomial", data = sleep)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.82148  -0.04746   0.00000   0.05811   2.41681
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -6.16387    3.59301  -1.716   0.0863 .
## brainweight          0.06018    0.03544   1.698   0.0895 .
## totalsleep           0.35985    0.20995   1.714   0.0865 .
## sleepexposureindex   4.42111    1.97540   2.238   0.0252 *
## predationindex      -3.36917    1.51823  -2.219   0.0265 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.310  on 50  degrees of freedom
## Residual deviance: 19.212  on 46  degrees of freedom
## AIC: 29.212
##
## Number of Fisher Scoring iterations: 11
```

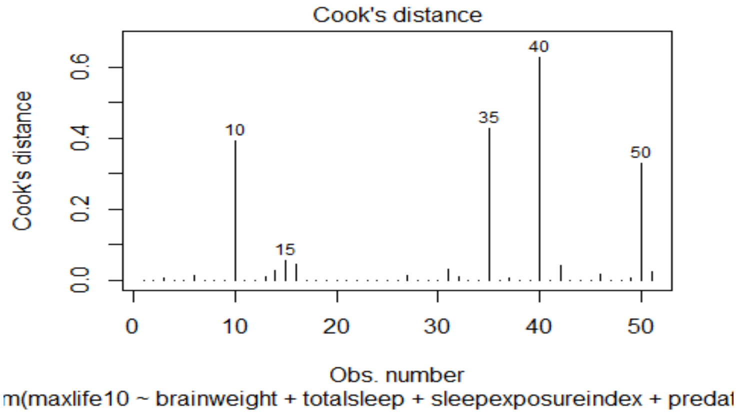
(b) All chosen predictors in the final model are statistically significant with p-values less than 0.1, meaning that these coefficients are significantly different from 0. Thus, it aids in predicting whether the maximum lifespan of a species will be at least 10 years. The goodness of fit test for the model has a p-value of 0.99, which indicates the model fit is reasonable.

We observe a few observations have cook's d relatively larger than others and residual plots does not show problematic issues.

```
hoslem.test(glm.sleep2$y, fitted(glm.sleep2), g=10)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
```

```
## data:  glm.sleep2$y, fitted(glm.sleep2)
## X-squared = 1.4406, df = 8, p-value = 0.9937
```





(c) The estimated OR for brainweight is 1.062 and it implies that the odds of a species having maximum lifespan at least 10 years is expected to increase by 1.062 (=exp(0.06)) times with one unit increase in brainweight. The estimated OR for totalsleep is 1.433 and it implies that the odds of a species having maximum lifespan at least 10 years is expected to increase by 1.433 (=exp(0.36)) with one unit increase in totalsleep. The odds ratio is estimated as 83.18 for sleepexposure so we can say for a one-unit increase in sleep exposure index, we expect to see an increase in the odds of a species having maximum lifespan at least 10 years by 83.18 (=exp(4.42)) times. The odds ratio for predation index is estimated as 0.034. Thus we expect the odds of a species having max lifespan at least 10 years change by 0.034 (=exp(-3.37)) multiplicative factor with one unit increase in predation index.

Estimated odds ratio is very large for sleep exposure index. We need to be careful to see if this result is presumable. But it is different issue and above is what we get from data. Also we see different result for the significance of variables from Exercise3 and 4. The reason is due to small sample size with relatively large number of parameters for the model fitted in Exercise3.

```
OR=exp(glm.sleep2$coefficients)
round(OR,3)

##        (Intercept)        brainweight         totalsleep sleepexposureindex 
##              0.002              1.062              1.433             83.188 
##     predationindex 
##              0.034 
```