Homework 4

Rudy Martinez, Brenda Parnin, Jose Fernandez

11/29/2020

Libraries

```
library(DescTools)
library(ResourceSelection)
```

Set Working Directory

```
setwd("/Users/rudymartinez/Desktop/MSDA/Fall 2020/STA 6443_Algorithms I/STAT-Algorithms-1/Week 14/Homework 4")
```

NOTE: Use significance level alpha = 0.1 in HW4

Exercise 1

The **liver** data set is a subset of the **ILPD** (Indian Liver Patient Dataset) data set. It contains the first 10 variables described on the UCI Machine Learning Repository and a **LiverPatient** variable (indicating whether or not the individual is a liver patient. People with active liver disease are coded as **LiverPatient**=1 and people without disease are coded **LiverPatient**=0) for adults in the data set. Adults here are defined to be individuals who are at least 18 years of age. It is possible that there will be different significant predictors of being a liver patient for adult females and adult males.

Read and View Dataset Structure

```r
liver = read.csv("liver.csv", header = TRUE)

str(liver)
```

```
## 'data.frame':    558 obs. of  10 variables:
##  $ Age         : int  65 62 62 58 72 46 26 29 55 57 ...
##  $ Gender      : chr  "Female" "Male" "Male" "Male" ...
##  $ TB          : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.7 0.6 ...
##  $ DB          : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.2 0.1 ...
##  $ Alkphos     : int  187 699 490 182 195 208 154 202 290 210 ...
##  $ Alamine     : int  16 64 60 14 27 19 16 14 53 51 ...
##  $ Aspartate   : int  18 100 68 20 59 14 12 11 58 59 ...
##  $ TP          : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 6.8 5.9 ...
##  $ ALB         : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 3.4 2.7 ...
##  $ LiverPatient: int  1 1 1 1 1 1 1 1 1 1 ...
```

Exercise 1.A

For only females in the data set, find and specify the best set of predictors via stepwise selection with AIC criteria for a logistic regression model predicting whether a female is a liver patient with active liver disease.

NOTE: Specifying the full model using "LiverPatient~., data=..." will give an error message (due to only one level of factor – Female – in the data, I guess so). Suggest typing all variables manually for the full model

```r
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")
```

Fit Logistic Regression Model (Female)

```r
step.model.1 = step(glm.null.F, scope = list(upper = glm.full.F),
                    direction="both",test="Chisq", trace = F)

summary(step.model.1)
```

Stepwise Selection with AIC Criteria

```
##
## 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
```

- **Significant Predictors:** Based on these results, `DB` and `Aspartate` are significant predictors as they have a p-value below the significance level of 0.1.

Exercise 1.B

Comment on the significance of parameter estimates under significance level **alpha=0.1**, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any issues with diagnostics by checking residual plots and cook's distance plot (**with cut-off 0.25**).

- **Significance of Parameter Estimates:** The `DB` predictor has a p-value of 0.0905, below the significance level of 0.1; therefore, there is a significant relationship between `DB` and whether a female is a liver patient with active liver disease. The `Aspartate` predictor has a p-value below of .0726, below the significance level of 0.1; therefore, there is a significant relationship between `Aspartate` and whether a female is a liver patient with active liver disease.

```
hoslem.test(step.model.1$y, fitted(step.model.1), g=10)
```

Hosmer-Lemeshow Test

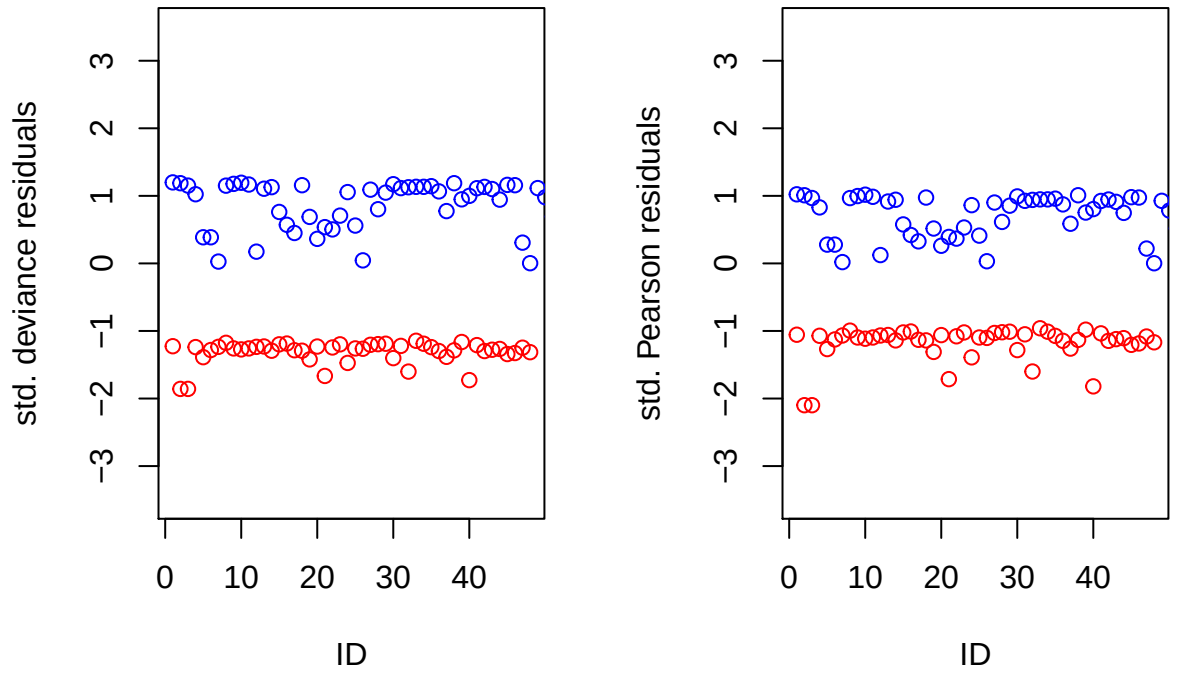
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  step.model.1$y, fitted(step.model.1)
## X-squared = 7.7535, df = 8, p-value = 0.4579
```

- **Goodness of Fit:** The Hosmer-Lemeshow Test yielded a p-value of 0.4579 which is above the significance level of 0.1. We **do not** reject the null; therefore, the **model is adequate**.

```
resid.d = residuals(step.model.1, type = "deviance")
resid.p = residuals(step.model.1, type = "pearson")
std.res.d = residuals(step.model.1, type = "deviance")/sqrt(1 - hatvalues(step.model.1))
std.res.p = residuals(step.model.1, type = "pearson")/sqrt(1 - hatvalues(step.model.1))

par(mfrow=c(1,2))
plot(std.res.d[step.model.1$model$LiverPatient==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. deviance residuals", xlab = "ID")
points(std.res.d[step.model.1$model$LiverPatient==1], col = "blue")
```

```r
plot(std.res.p[step.model.1$model$LiverPatient==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. Pearson residuals", xlab = "ID")
points(std.res.p[step.model.1$model$LiverPatient==1], col = "blue")
```



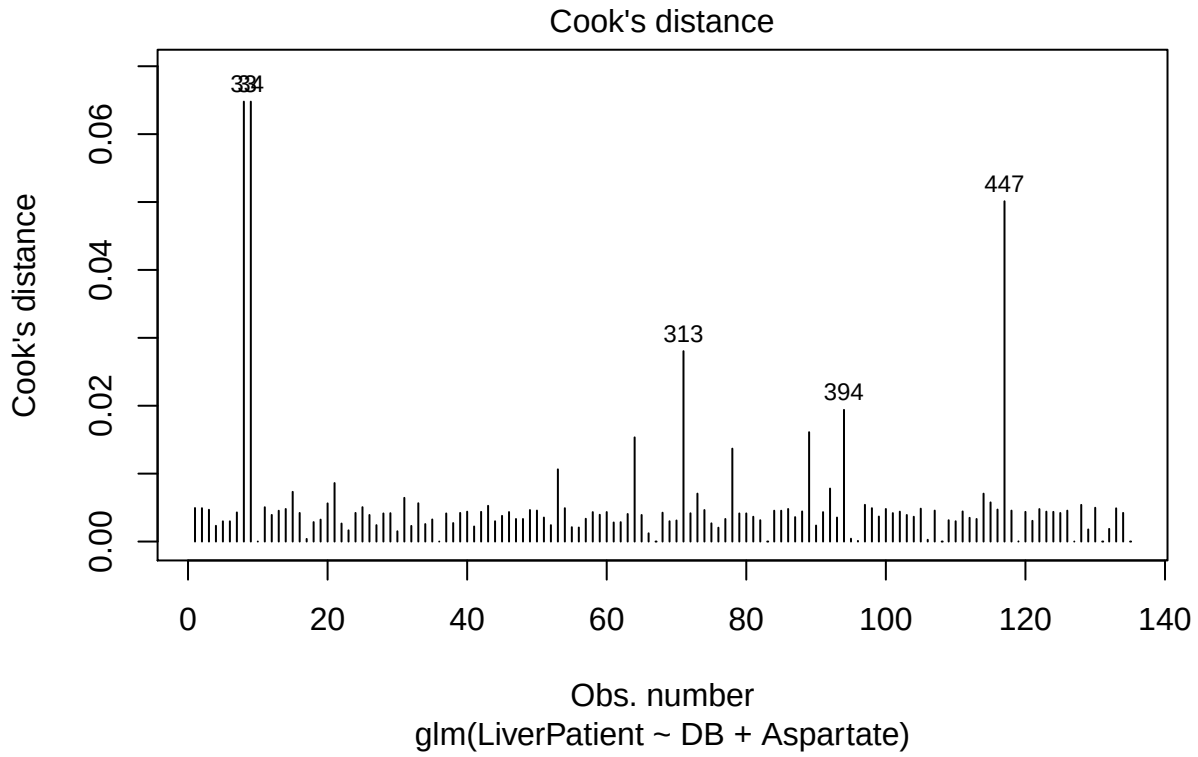
Residual Plots

- **Observation:** There appears to be a parallel pattern in the residuals plots. This is due to similar estimated probabilities for all observations. To explain, both Pearson and Deviance residuals are based on (Y - P_hat) and if the p_hats are similar for all observations, a parallel pattern presents itself. Therefore, the parallel pattern above is due to data feature, not due to violation of assumptions. Additionally, the plotted points fall within the range of 0 to 1 (Blue) and -1 to -2 (Red).

Because there are no points with very large values, the Bernoulli assumption is valid. Moreover, because there is not a systematic pattern in the plot;

therefore, it does not violate the linearity assumption.

```
plot(step.model.1, which = 4, id.n = 5)
```



Cook's Distance (Influence Diagnostics)

```
inf.id.1 = which(cooks.distance(step.model.1)>0.25)
inf.id.1
```

```
## named integer(0)
```

- **Issues Identified in Model Diagnostics:** There are no observations with a Cook's distance larger than 0.25.

Exercise 1.C

Interpret relationships between predictors in the final model and the odds of an adult female being a liver patient. (based on estimated Odds Ratio).

```
summary(step.model.1)
```

Final Model

```
##
## 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
```

Final Model: Log(p/1-p) = -0.32480 + 0.94479 * `DB` + 0.01106 * `Aspartate`

```r
round(exp(step.model.1$coefficients),3)
```

Odds Ratio

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

Odds Ratio Interpretation:

- The odds of an adult female being a liver patient with active liver disease increase by a factor of $\exp(0.94479) = 2.572$ with a one unit increase in `DB` when `Aspartate` is held constant.
- The odds of an adult female being a liver patient with active liver disease increase by a factor of $\exp(0.01106) = 1.011$ with a one unit increase in `Aspartate` when `DB` is held constant.

Thus, an adult female with high levels of `DB` or Direct Bilirubin and `Aspartate` or Aspartate Aminotransferase is more likely to be a liver patient with active liver disease.

Exercise 2

Repeat exercise 1 for males. In addition to the previous questions, also d) comment on how the models for adult females and adult males differ. Use significance level **alpha=0.1**

Exercise 2.A

For only males in the data set, find and specify the best set of predictors via stepwise selection with AIC criteria for a logistic regression model predicting whether a male is a liver patient with active liver disease.

```r
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")
```

Fit Logistic Regression Model (Male)

```
step.model.2 = step(glm.null.M, scope = list(upper = glm.full.M),
                    direction="both",test="Chisq", trace = F)

summary(step.model.2)
```

Stepwise Selection with AIC Criteria

```
##
## Call:
## glm(formula = LiverPatient ~ DB + Alamine + Age + Alkphos, family = "binomial",
##     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
```

- **Significant Predictors:** Based on these results, `DB`, `Alamine`, `Age`, and `Alkphos` are significant predictors as they have a p-value below the significance level of 0.1.

Exercise 2.B

Comment on the significance of parameter estimates under significance level **alpha=0.1**, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any issues with diagnostics by checking residual plots and cook's distance plot (**with cut-off 0.25**).

- **Significance of Parameter Estimates:** The `DB`, `Alamine`, `Age`, and `Alkphos` predictors have p-values of 0.00360, 0.00197, 0.01087, and 0.09992 respectively, below the significance level of 0.1; therefore, there is a significant relationship between `DB`, `Alamine`, `Age`, and `Alkphos` and whether a male is a liver patient with active liver disease.

```
hoslem.test(step.model.2$y, fitted(step.model.2), g=10)
```

Hosmer-Lemeshow Test

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  step.model.2$y, fitted(step.model.2)
## X-squared = 7.043, df = 8, p-value = 0.532
```

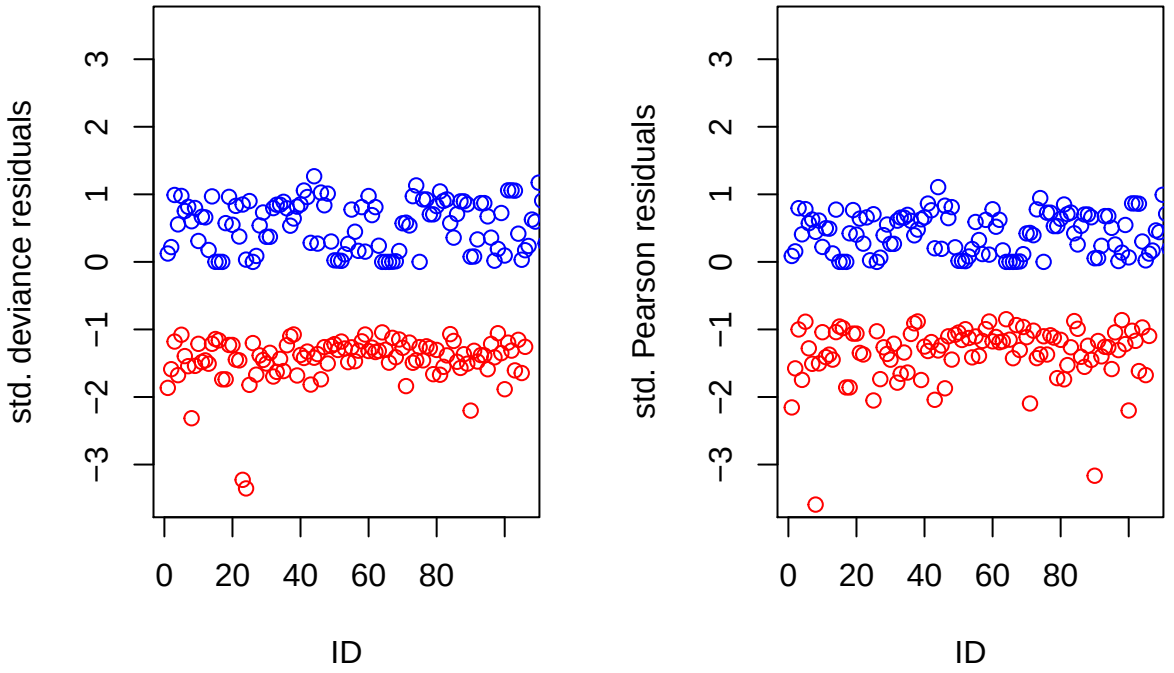
- **Goodness of Fit:** The Hosmer-Lemeshow Test yielded a p-value of 0.532 which is above the significance level of 0.1. We **do not** reject the null; therefore, the **model is adequate**.

```
resid.d = residuals(step.model.2, type = "deviance")
resid.p = residuals(step.model.2, type = "pearson")
std.res.d = residuals(step.model.2, type = "deviance")/sqrt(1 - hatvalues(step.model.2))
std.res.p = residuals(step.model.2, type = "pearson")/sqrt(1 - hatvalues(step.model.2))

par(mfrow=c(1,2))
```

```r
plot(std.res.d[step.model.2$model$LiverPatient==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. deviance residuals", xlab = "ID")
points(std.res.d[step.model.2$model$LiverPatient==1], col = "blue")

plot(std.res.p[step.model.2$model$LiverPatient==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. Pearson residuals", xlab = "ID")
points(std.res.p[step.model.2$model$LiverPatient==1], col = "blue")
```



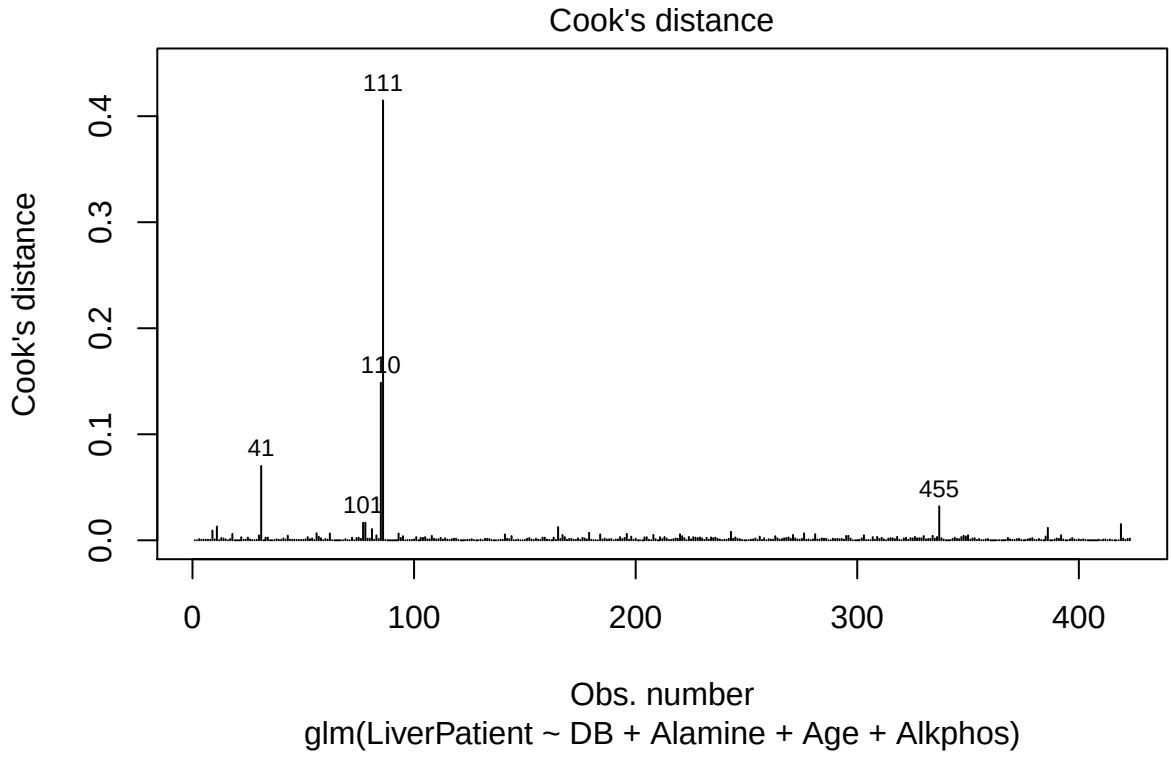
Residual Plots

- **Observation:** There appears to be a parallel pattern in the residuals plots. This is due to similar estimated probabilities for all observations. To explain, both Pearson and Deviance residuals are based on (Y - P_hat) and if the p_hats are similar for all observations, a parallel pattern

presents itself. Therefore, the parallel pattern above is due to data feature, not due to violation of assumptions. Additionally, the majority of the plotted points fall within the range of 0 to 1 (Blue) and -1 to -2 (Red). It's worth noting that there are a set of Red points outside of the -1 to -2 range.

Because there are no points with very large values, the Bernoulli assumption is valid. Moreover, because there is not a systematic pattern in the plot; therefore, it does not violate the linearity assumption.

```
plot(step.model.2, which = 4, id.n = 5)
```



Cook's Distance (Influence Diagnostics)

```r
inf.id.2 = which(cooks.distance(step.model.2)>0.25)
inf.id.2
```

```
## 111
##  86
```

- **Issues Identified in Model Diagnostics:** Observation **111** and **86** have a Cook's distance larger than 0.25.

```r
glm.liver.final.2 = glm(LiverPatient ~ DB + Alamine + Age + Alkphos, data = liverM[-inf.id.2, ], family = "binomial")
```

Refitted Model without Observations 111 and 86:

Exercise 2.C

Interpret relationships between predictors in the final model and the odds of an adult male being a liver patient. (based on estimated Odds Ratio).

```r
summary(glm.liver.final.2)
```

Final Model

```
##
## Call:
## glm(formula = LiverPatient ~ DB + Alamine + Age + Alkphos, family = "binomial",
##     data = liverM[-inf.id.2, ])
##
## 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
```

Final Model: : Log(p/1-p) = -1.902754 + 0.573104 * `DB` + 0.015850 * `Alamine` + 0.020418 * `Age` + 0.003744 * `Alkphos`

```
round(exp(glm.liver.final.2$coefficients),3)
```

Odds Ratio

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

Odds Ratio Interpretation:

- The odds of an adult male being a liver patient with active liver disease increase by a factor of exp(0.573104) = 1.774 with a one unit increase in `DB` when `Alamine`, `Age`, and `Alkphos` are held constant.
- The odds of an adult male being a liver patient with active liver disease increase by a factor of exp(0.015850) = 1.016 with a one unit increase in `Alamine` when `DB`, `Age`, and `Alkphos` are held constant.
- The odds of an adult male being a liver patient with active liver disease increase by a factor of exp(0.020418) = 1.021 with a one unit increase in `Age` when `DB`, `Alamine`, and `Alkphos` are held constant.

- The odds of an adult male being a liver patient with active liver disease increase by a factor of exp(0.003744) = 1.004 with a one unit increase in `Alkphos` when `DB`, `Alamine`, and `Age` are held constant.

Thus, an adult male with high levels of `DB` or Direct Bilirubin, `Alamine` or Alamine Aminotransferase, `Alkphos` or Alkaline Phosphotase, and an older `Age` is more likely to be a liver patient with active liver disease.

Exercise 2.D

Comment on how the models for adult females and adult males differ.

Adult females have fewer predictors (`DB` and `Aspartate`) that are significant and increase the odds of being a liver patient with active liver disease. Males have more predictors (`DB`, `Alamine`, `Age`, and `Alkphos`) that are significant and increase the odds of being a liver patient with active liver disease.

Exercise 3

Use the **sleep_new data** set which originates from http://lib.stat.cmu.edu/datasets/sleep. **maxlife10** is 1 if the species maximum life span is less than 10 years and 0 if its maximum life span is greater than or equal to 10 years. Consider finding the best logistic model for predicting the probability that a species' maximum lifespan will be at least 10 years. Consider all 6 variables as candidates (do not include species) and two index variables of them are categorical in nature. **Treat two index variables as categorical variables** (e.g. ignore the fact that they are ordinal). Use significance level **alpha=0.1**

Read and View Dataset Structure

```
sleep = read.csv("sleep_new.csv", header = TRUE)

str(sleep)
```

```
## 'data.frame':    51 obs. of  8 variables:
##  $ species           : chr  "African" "African" "Arctic F" "Asian el" ...
##  $ bodyweight        : num  6654 1 3.38 2547 10.55 ...
##  $ brainweight       : num  5712 6.6 44.5 4603 179.5 ...
##  $ totalsleep        : num  3.3 8.3 12.5 3.9 9.8 19.7 6.2 14.5 9.7 12.5 ...
##  $ gestationtime     : num  645 42 60 624 180 35 392 63 230 112 ...
##  $ predationindex    : int  3 3 1 3 4 1 4 1 1 5 ...
##  $ sleepexposureindex: int  5 1 1 5 4 1 5 2 1 4 ...
##  $ maxlife10         : int  1 0 1 1 1 1 1 1 1 0 ...
```

Exercise 3.A

First find and specify the best set of predictors via stepwise selection with AIC criteria.

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

glm.full.sleep1 = glm(maxlife10 ~ bodyweight + brainweight + totalsleep + gestationtime + as.factor(predationindex) + as.factor(sleepexposureindex),

                      data = sleep, family = "binomial")
```

Fit Logistic Regression Model

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

summary(step.sleep1)
```

Stepwise Selection with AIC Criteria

```
##
## Call:
## glm(formula = maxlife10 ~ brainweight + totalsleep + as.factor(sleepexposureindex) +
##     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
```

- **Significant Predictors:** Based on these results, `sleepexposureindex` is a significant predictor as it has a p-value below the significance level of 0.1.

Exercise 3.B

Comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any issues with diagnostics by checking residual plots and cook's distance plot. **Do not** remove influential points but just make comments on suspicious observations.

- **Significance of Parameter Estimates:** `sleepexposureindex` Level 1,3,4, and 5 have no effect on the probability of having an event. Only `sleepexposureindex` Level 2 has a significantly different probability of having an event. Even though this is the case, `sleepexposureindex` is a **significant** predictor.

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

Hosmer-Lemeshow Test

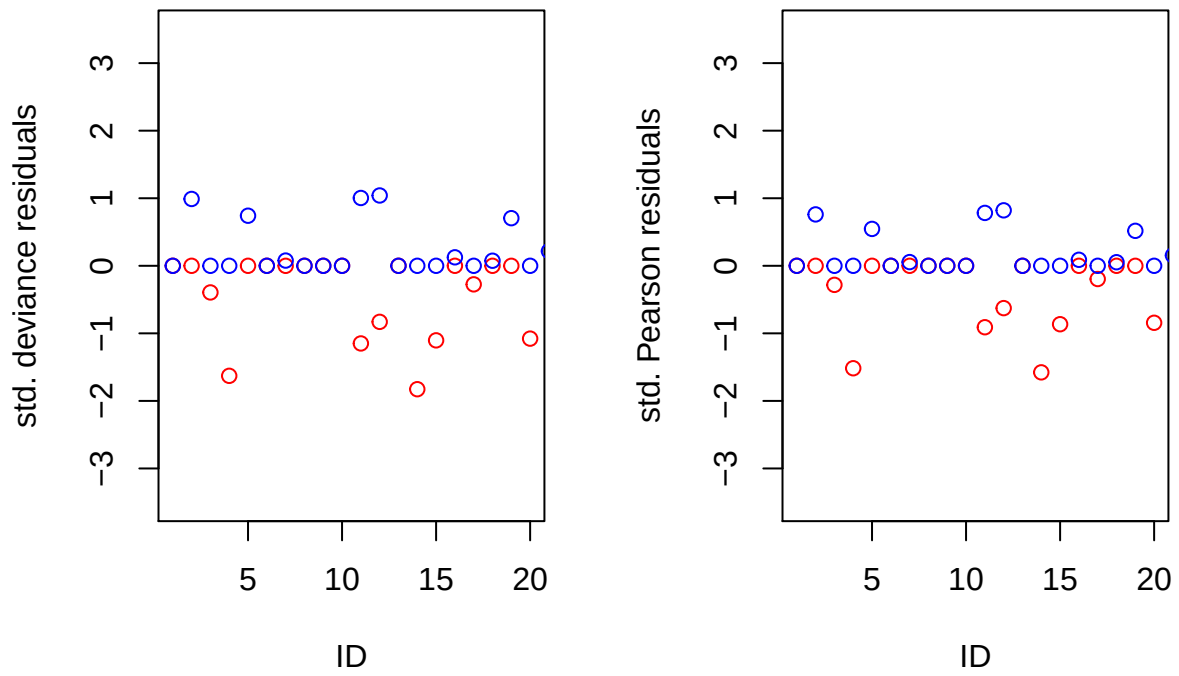
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  step.sleep1$y, fitted(step.sleep1)
## X-squared = 7.0397, df = 8, p-value = 0.5324
```

- **Goodness of Fit:** The Hosmer-Lemeshow Test yielded a p-value of 0.5324 which is above the significance level of 0.1. We **do not** reject the null; therefore, the **model is adequate**.

```
resid.d = residuals(step.sleep1, type = "deviance")
resid.p = residuals(step.sleep1, type = "pearson")
std.res.d = residuals(step.sleep1, type = "deviance")/sqrt(1 - hatvalues(step.sleep1))
std.res.p = residuals(step.sleep1, type = "pearson")/sqrt(1 - hatvalues(step.sleep1))

par(mfrow=c(1,2))
plot(std.res.d[step.sleep1$model$maxlife10==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. deviance residuals", xlab = "ID")
points(std.res.d[step.sleep1$model$maxlife10==1], col = "blue")

plot(std.res.p[step.sleep1$model$maxlife10==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. Pearson residuals", xlab = "ID")
points(std.res.p[step.sleep1$model$maxlife10==1], col = "blue")
```

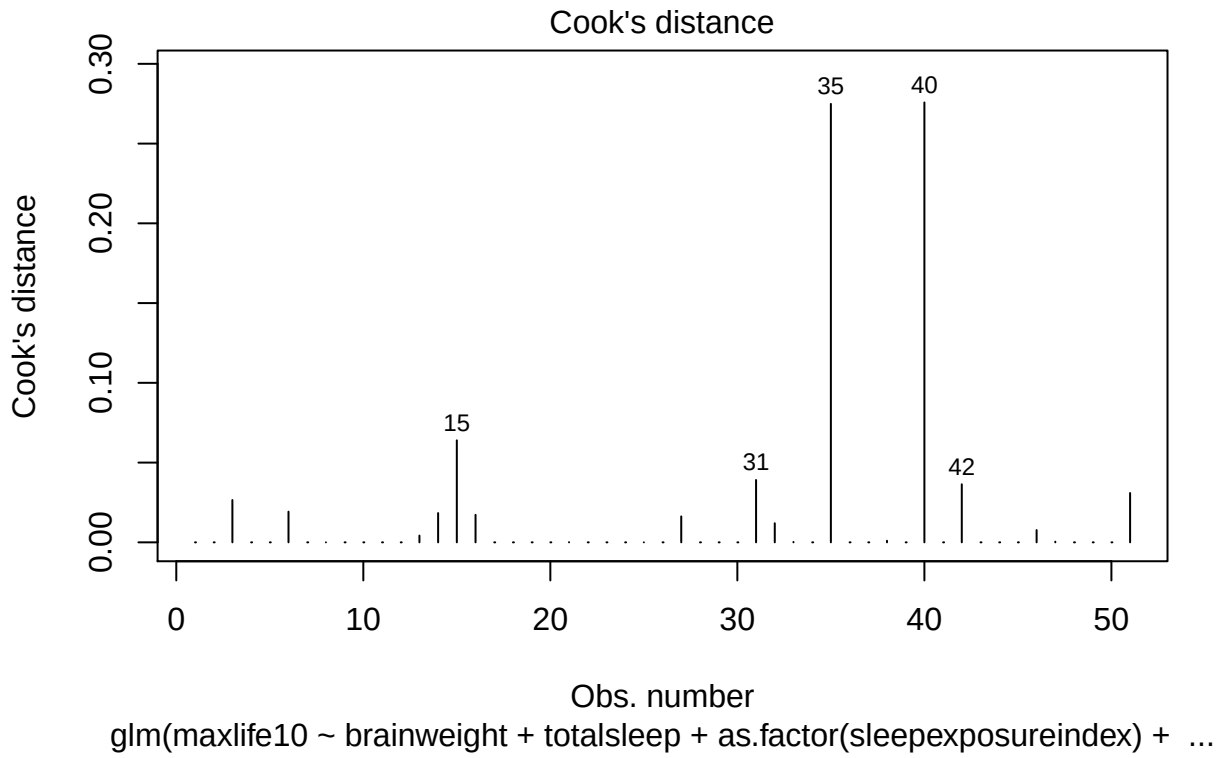


Residual Plots

- **Observation:** All of the Blue plotted points fall within the range of 0 to 1; however, the Red plotted points fall within the range of 0 to -2. It appears that there is some overlap in the plotted points.

Because there are no points with very large values, the Bernoulli assumption is valid. Moreover, because there is not a systematic pattern in the plot; therefore, it does not violate the linearity assumption.

```
plot(step.sleep1, which = 4, id.n = 5)
```



Cook's Distance (Influence Diagnostics)

```
inf.id.3 = which(cooks.distance(step.sleep1)>0.25)
inf.id.3
```

```
## 35 40
## 35 40
```

- **Issues Identified in Model Diagnostics:** Observation **35** and **40** have a Cook's Distance larger than 0.25.

```r
glm.sleep.final.1 = glm(maxlife10 ~ brainweight + totalsleep + as.factor(predationindex) + as.factor(sleepexposureindex),
                        data = sleep[-inf.id.3], family = "binomial")
```

Refitted Model without Observations 35 and 40:

Exercise 3.C

Interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years.

```r
summary(glm.sleep.final.1)
```

Final Model

```
##
## Call:
## glm(formula = maxlife10 ~ brainweight + totalsleep + as.factor(predationindex) +
##     as.factor(sleepexposureindex), family = "binomial", data = sleep[-inf.id.3])
##
## 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(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
## 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
## ---
## 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
```

```
round(exp(glm.sleep.final.1$coefficients),3)
```

Odds Ratio

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

Odds Ratio Interpretation:

- odds(`sleepexposureindex` = 2)/odds(`sleepexposureindex` = 1) = exp(4.998e+00) = 1.480500e+02
- There is no need to interpret the insignificant levels as they imply zero coefficients (with large p-values). The odds that a species' maximum lifespan will be at least 10 years is exp(4.998e+00) = 1.480500e+02 times for the `sleepexposureindex` Level 2 group compared other groups.

Thus, animals that sleep in the second-best (Level 2) well-protected den have a higher probability of achieving a maximum lifespan of at least 10 years.

Exercise 4

The index variables in the data set are ordinal, meaning they are categorical and they have a natural ordering. If we treat an index variable as a continuous variable, this will imply a linear change as the index changes. Repeat Exercise 3 by **treating two index variables as continuous variables**. Use significance level **alpha=0.1**

```
glm.null.sleep2 = glm(maxlife10 ~ 1, data = sleep, family = "binomial")

glm.full.sleep2 = glm(maxlife10 ~ bodyweight + brainweight + totalsleep + gestationtime + predationindex + sleepexposureindex, data = slee
```

Fit Logistic Regression Model

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

summary(step.sleep2)
```

Stepwise Selection with AIC Criteria

```
##
## 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
```

- **Significant Predictors:** Based on these results, `brainweight`, `totalsleep`, `sleepexposureindex`, and `predationindex` are a significant predictors as they have a p-value below the significance level of 0.1

Exercise 4.B

Comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any issues with diagnostics by checking residual plots and cook's distance plot. **Do not** remove influential points but just make comments on suspicious observations.

- **Significance of Parameter Estimates:** The `brainweight`, `totalsleep`, `sleepexposureindex`, and `predationindex` predictors have p-values of 0.0895, 0.0865, 0.0252, and 0.0265 respectively, below the significance level of 0.1; therefore, there is a significant relationship between `brainweight`, `totalsleep`, `sleepexposureindex`, and `predationindex` and whether an animal species' maximum lifespan will be at least 10 years.

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

Hosmer-Lemeshow Test

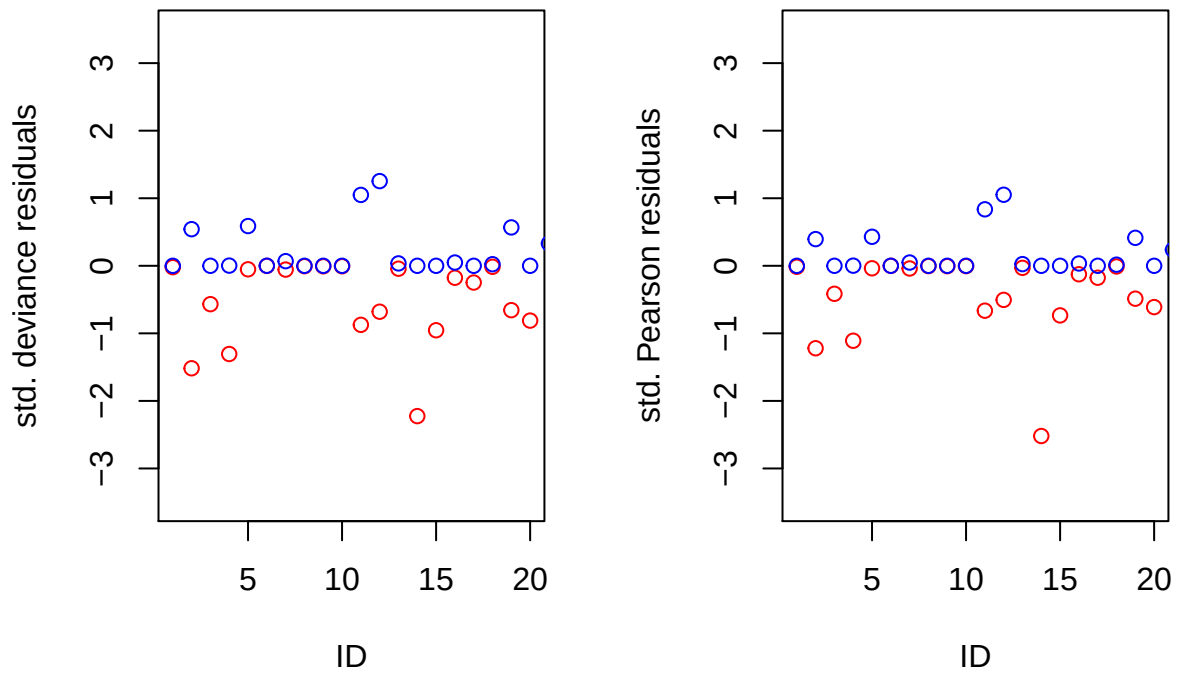
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  step.sleep2$y, fitted(step.sleep2)
## X-squared = 1.4406, df = 8, p-value = 0.9937
```

- **Goodness of Fit:** The Hosmer-Lemeshow Test yielded a p-value of 0.9937 which is above the significance level of 0.1. We **do not** reject the null; therefore, the **model is adequate**.

```
resid.d = residuals(step.sleep2, type = "deviance")
resid.p = residuals(step.sleep2, type = "pearson")
std.res.d = residuals(step.sleep2, type = "deviance")/sqrt(1 - hatvalues(step.sleep2))
std.res.p = residuals(step.sleep2, type = "pearson")/sqrt(1 - hatvalues(step.sleep2))

par(mfrow=c(1,2))
plot(std.res.d[step.sleep2$model$maxlife10==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. deviance residuals", xlab = "ID")
points(std.res.d[step.sleep2$model$maxlife10==1], col = "blue")

plot(std.res.p[step.sleep2$model$maxlife10==0], col = "red",
     ylim = c(-3.5,3.5), ylab = "std. Pearson residuals", xlab = "ID")
points(std.res.p[step.sleep2$model$maxlife10==1], col = "blue")
```

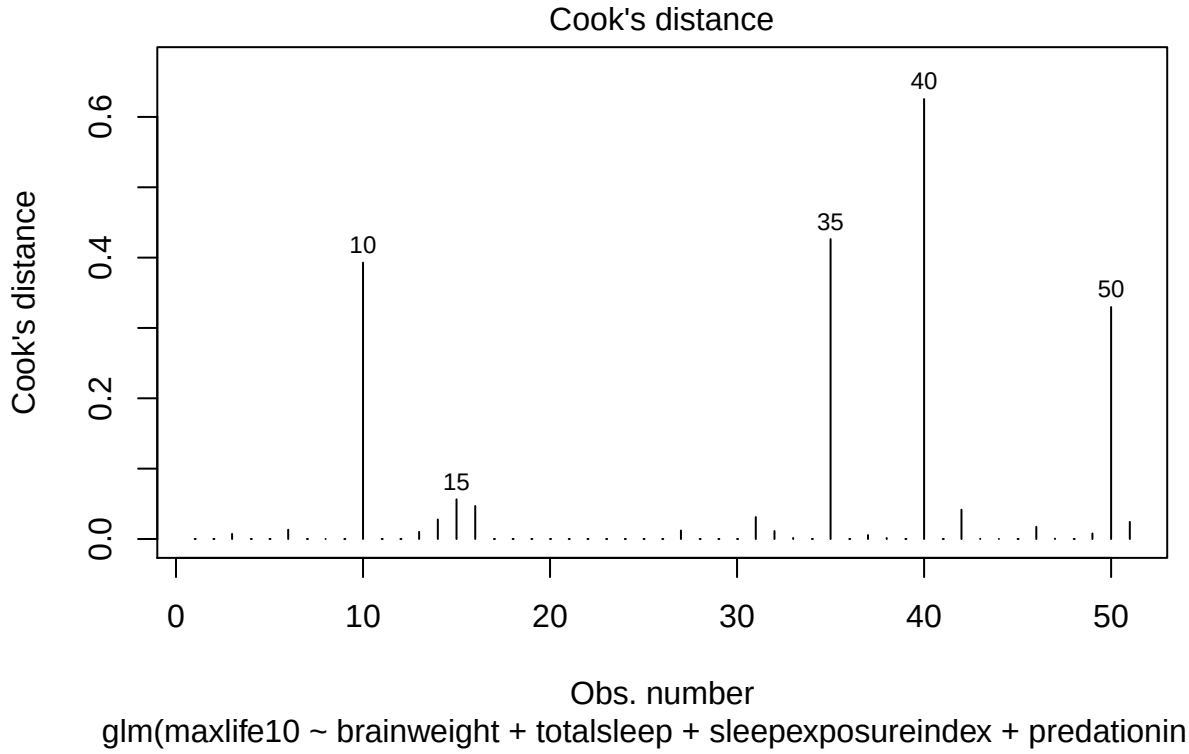


Residual Plots

- **Observation:** All of the Blue plotted points fall within the range of 0 to 1; however, the majority of the Red plotted points fall within the range of 0 to -2. It appears that there is some overlap in the plotted points, and there is also a Red plotted point outside the range of 0 to -2.

Because there are no points with very large values, the Bernoulli assumption is valid. Moreover, because there is not a systematic pattern in the plot; therefore, it does not violate the linearity assumption.

```
plot(step.sleep2, which = 4, id.n = 5)
```



Cook's Distance (Influence Diagnostics)

```
inf.id.4 = which(cooks.distance(step.sleep2)>0.25)
inf.id.4
```

```
## 10 35 40 50
## 10 35 40 50
```

- **Issues Identified in Model Diagnostics:** Observation **10**, **35**, **40**, and **50** have a Cook's Distance larger than 0.25.

```r
glm.sleep.final.2 = glm(maxlife10 ~ brainweight + totalsleep + predationindex + sleepexposureindex,
                        data = sleep[-inf.id.4], family = "binomial")
```

Refitted Model without Observations 10, 35, 40, and 50

Exercise 4.C

Interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years.

```r
summary(glm.sleep.final.2)
```

Final Model

```
##
## Call:
## glm(formula = maxlife10 ~ brainweight + totalsleep + predationindex +
##     sleepexposureindex, family = "binomial", data = sleep[-inf.id.4])
##
## 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 .
## predationindex     -3.36917    1.51823  -2.219   0.0265 *
## sleepexposureindex  4.42111    1.97540   2.238   0.0252 *
## ---
## 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
```

Final Model: Log(p/1-p) = -6.16387 + 0.06018 * `brainweight` + 0.35985 * `totalsleep` + 4.42111 * `sleepexposureindex` + -3.36917 * `predationindex`

```
round(exp(glm.sleep.final.2$coefficients),3)
```

Odds Ratio

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

Odds Ratio Interpretation:

- The odds that a species' maximum lifespan will be at least 10 years increase by a factor of exp(0.06018) = 1.062 with a one unit increase in `brainweight` when `totalsleep`, `sleepexposureindex`, and `predationindex` are held constant.
- The odds that a species' maximum lifespan will be at least 10 years increase by a factor of exp(0.35985) = 1.433 with a one unit decrease in `totalsleep` when `brainweight`, `sleepexposureindex`, and `predationindex` are held constant.
- The odds that a species' maximum lifespan will be at least 10 years decrease by a factor of exp(-3.36917) = 0.034 with a one unit increase in `predationindex` when `totalsleep`, `brainweight`, and `sleepexposureindex` are held constant.
- The odds that a species' maximum lifespan will be at least 10 years increase by a factor of exp(4.42111) = 83.188 with a one unit increase in `sleepexposureindex` when `totalsleep`, `brainweight`, and `predationindex` are held constant.

Thus, a species of an animal that has higher `brainweight`, has higher `totalsleep`, has a lower `predationindex`, and has a lower `sleepexposureindex` is more likely to have a lifespan of at least 10 years. Simply put, an animal species with a heavier brain that gets more sleep, is least likely to be preyed upon, and sleeps in a less exposed area has a higher probability of having a lifespan of at least 10 years.