HW4

AUTHOR
Collin Real (yhi267), Seth Harris (dmp903)

Import libraries.

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.3     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.3     ✔ tibble    3.2.1
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(ResourceSelection)
```

```
ResourceSelection 0.3-6      2023-06-27
```

```
library(DescTools)
```

Set path.

```
setwd("/Users/c2cypher/codebase/msda/msda-grad-school/sta-6443-902-data_analytics_algo
```

To complete this assignment, you will need to download liver.csv and sleep.csv. We 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.

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.

Importing Data

```
liverdata<- read.csv("liver-1.csv", header = TRUE)
str(liverdata)
```

```
'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 ...
```

Running Fit Logistic Regression Model (Female)

```
liverFemale = liverdata[which(liverdata$Gender == "Female"),]

glmnullFemale = glm(LiverPatient ~ 1, data = liverFemale, family = "binomial")

glmfullFemale = glm(LiverPatient ~ Age + TB + DB +Alkphos + Alamine + Aspartate + TP +
```

Stepwise Selection with AIC Criteria

```
stepwiseselectionfemale = step(glmnullFemale, scope = list(upper = glmfullFemale),
                     direction="both",test="Chisq", trace = F)

summary(stepwiseselectionfemale)
```

```
Call:
glm(formula = LiverPatient ~ DB + Aspartate, family = "binomial",
    data = liverFemale)

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
```

Comments Overall, it is clear that our variables **DB and Aspartate** are significant predictors due to the fact that their p-values of DB (0.0905) and Aspartate (0.0726) are lower than the significance level of 0.10. In addition, the Final Model (Female Only), which is Log(LiverPatient) = -0.32480 + 0.94479 * DB + 0.01106 * Aspartate might include insignificant variables.

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).

Comments When looking at the significance of parameter estimates under significance level alpha= 0.1, both our variables DB and Aspartate contain p-values of 0.0905 and 0.0726, which is below the significance level of 0.10. Therefore, as a result we can conclude that they are both significant predictors for female liver patients. Both of these specific variables (DB and Aspartate) have a increased chance of a relationship between female liver patients and active liver disease.

Calculating Goodness of Fit Using Hosmer-Lemeshow's Test

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

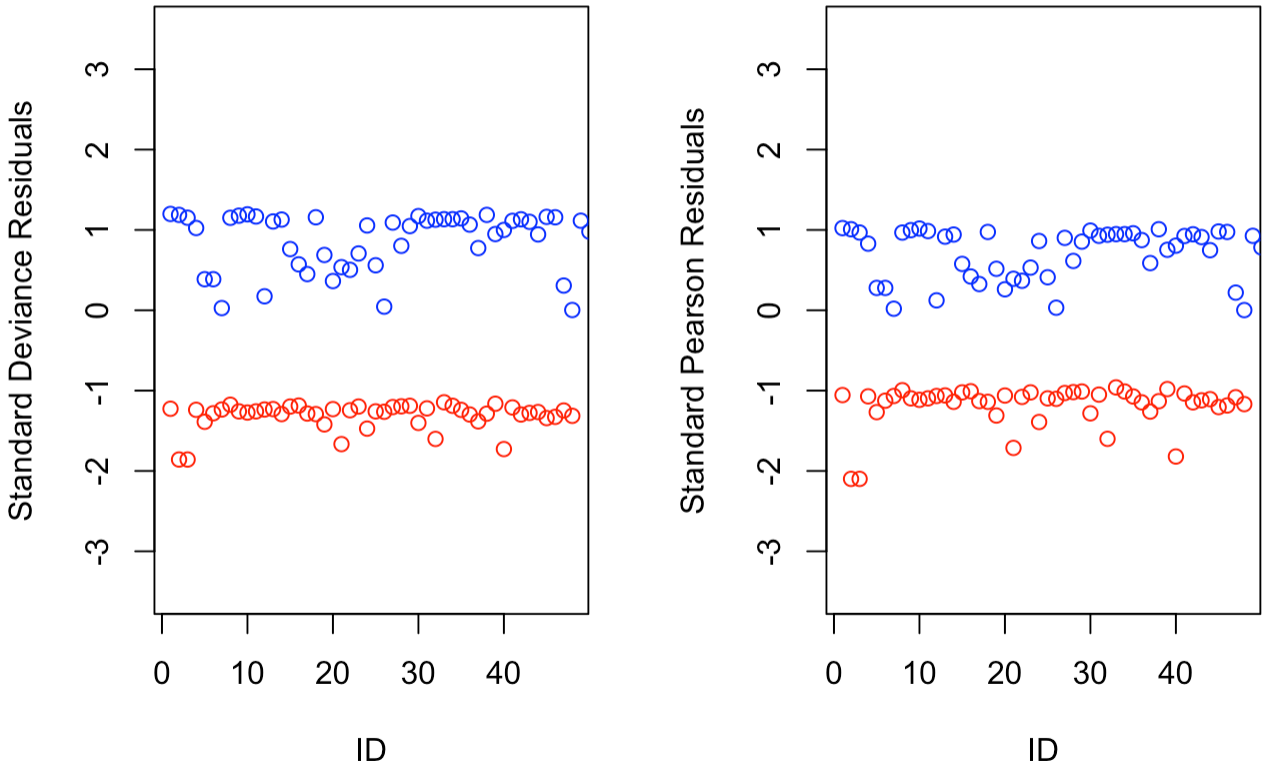
```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepwiseselectionfemale$y, fitted(stepwiseselectionfemale)
X-squared = 7.7535, df = 8, p-value = 0.4579
```

Comments on Goodness of Fit H0: Our model is adequate H1: Our model is not adequate As we can see shown above, the Hosmer-Lemeshow's Goodness of Fit Test contained a p-value of 0.4579, which is clearly above the significance level of 0.10, therefore we cannot reject our null hypothesis (H0), which means the model fits well and is considered to be adequate.

Checking Residual Plots

```
residualdeviance = residuals(stepwiseselectionfemale, type = "deviance")
residualpearson = residuals(stepwiseselectionfemale, type = "pearson")
standardresidualdeviance = residuals(stepwiseselectionfemale, type = "deviance")/sqrt(
standardresidualpearson = residuals(stepwiseselectionfemale, type = "pearson")/sqrt(1

par(mfrow=c(1,2))
plot(standardresidualdeviance[stepwiseselectionfemale$model$LiverPatient==0], col = "r
     ylim = c(-3.5,3.5), ylab = "Standard Deviance Residuals", xlab = "ID")
points(standardresidualdeviance[stepwiseselectionfemale$model$LiverPatient==1], col =

plot(standardresidualpearson[stepwiseselectionfemale$model$LiverPatient==0], col = "re
     ylim = c(-3.5,3.5), ylab = "Standard Pearson Residuals", xlab = "ID")
```

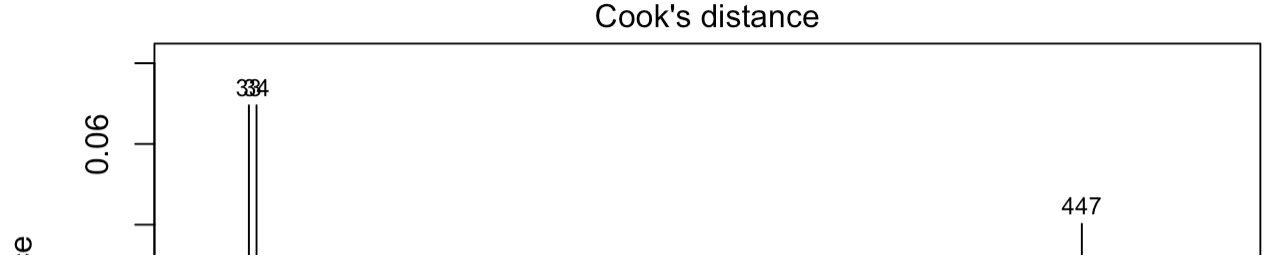
```
points(standardresidualpearson[stepwiseselectionfemale$model$LiverPatient==1], col = "
```

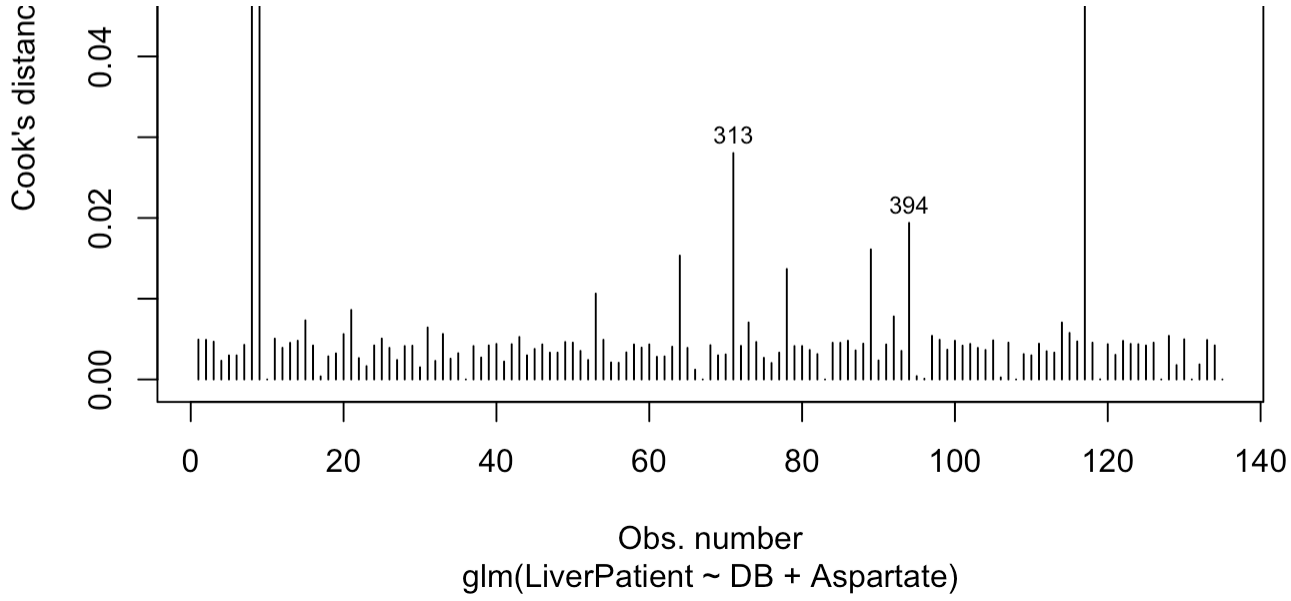


Comments The residual plots follow a similar pattern in both the Standard Deviance Residuals and Standard Pearson Residuals. As we can see above, the red line does not go away from linearity and there is no large outstanding values it does not violate any assumptions. As a result, because there really are not any points with extremely big values, the Bernoulli assumption is valid. In addition, because there is no systematic pattern in the plot, we can conclude that the homoscedasticity assumption is true.

Using Cook's Distance Plot

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






```
influenceddiagnostics = which(cooks.distance(stepwiseselectionfemale)>0.25)
influenceddiagnostics
```

```
named integer(0)
```

Comments As a result from Cook's Distance Plot, we can see there are no influential points (observations) made with a Cook's Distance Plot greater than 0.25.

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). NOTE: stepwise selection with AIC criteria can be performed by default step() function in R. **Final Model**

```
summary(stepwiseselectionfemale)
```

```
Call:
glm(formula = LiverPatient ~ DB + Aspartate, family = "binomial",
    data = liverFemale)

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
```

```
AIC: 100.27

Number of Fisher Scoring iterations: 7
```

Comments The final model would be **log(p/1-p) = -0.32480 + 0.94479 * DB + 0.01106 * Aspartate**

Calculating Odds Ratio

```
round(exp(stepwiseselectionfemale$coefficients),3)
```

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

Interpretation of Odds Ratio As we can see shown above, we make the conclusion that the probability of female being a liver patient with active liver disease can be increased by a factor of exp(0.94479) = 2.572 with a one unit increase in DB when our variable Aspartate is held constant. On the other hand, the probability of female being a liver patient with active liver disease increases by a factor of exp(0.01106) = 1.011 with a one unit increase in Aspartate when our variable DB is held constant. As a result, a female with high levels of our variables Direct Bilirubin (DB) and Aspartate Aminotransferase (Aspartate) 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** NOTE: You will get an error message “glm.fit: fitted probabilities numerically 0 or 1 occurred” for this run. Ignore this and use the result for the interpretation. I will explain what this error means in the class.

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 female is a liver patient.

Running Fit Logistic Regression Model (Male)

```
liverMale = liverdata[which(liverdata$Gender == "Male"),]

glmnullMale= glm(LiverPatient ~ 1, data = liverMale, family = "binomial")

glmfullMale = glm(LiverPatient ~ Age + TB + DB +Alkphos + Alamine + Aspartate + TP + A
```

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

Stepwise Selection with AIC Criteria

```
stepwiseselectionmale = step(glmnullMale, scope = list(upper = glmfullMale),
                    direction="both",test="Chisq", trace = F)
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
summary(stepwiseselectionmale)
```

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

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
```

Comments Overall, it is clear that our variables **DB, Alamine, Age, and Alkphos** are significant predictors due to the fact that their p-values of DB (0.00360), Alamine (0.00197), Age (0.01087), and Alkphos (0.09992) are lower than the significance level of 0.10. In addition, the Final Model (Male Only), which is Log(LiverPatient) = -1.476570 + 0.512503 * DB + 0.016218 * Alamin + 0.020616 * Age + 0.001740 * Alkphos might include insignificant variables.

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).

Comments When looking at the significance of parameter estimates under significance level alpha= 0.1, all our variables DB (0.00360), Alamine (0.00197), Age (0.01087), and Alkphos (0.09992), are

below the significance level of 0.10. Therefore, as a result we can conclude that they are both significant predictors for male liver patients. Both of these specific variables (DB, Alamine, Age, and Alkphos) have a increased chance of a relationship between male liver patients and active liver disease.

Calculating Goodness of Fit Using Hosmer-Lemeshow's Test

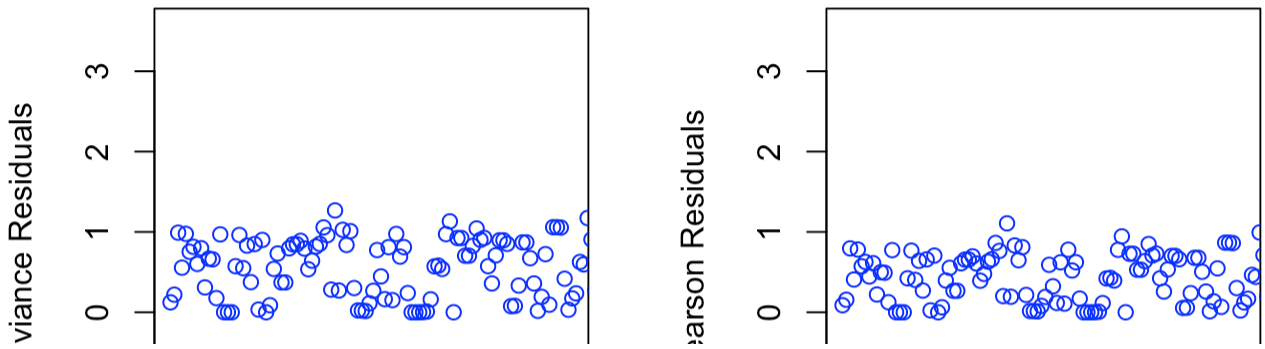
```
hoslem.test(stepwiseselectionmale$y, fitted(stepwiseselectionmale), g=10)
```

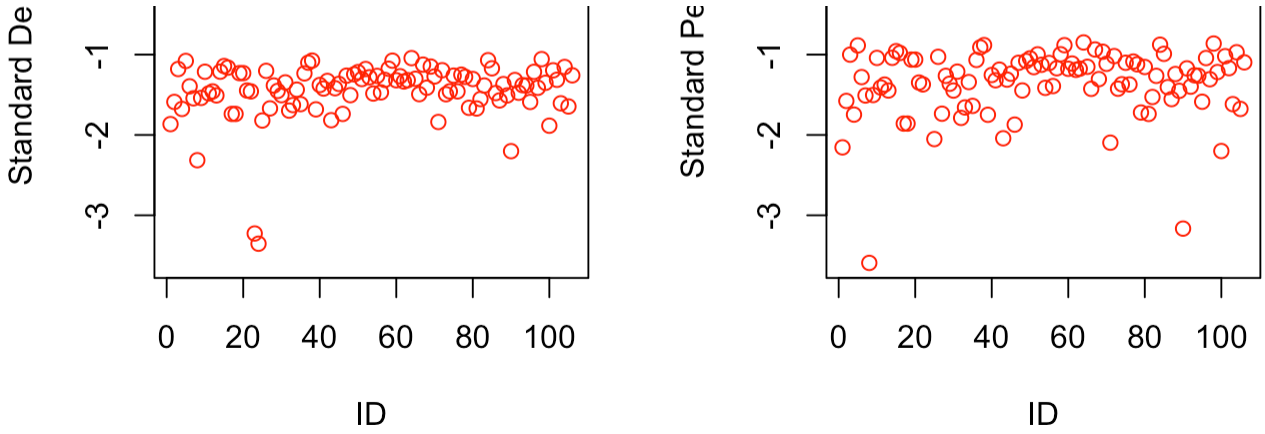
```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepwiseselectionmale$y, fitted(stepwiseselectionmale)
X-squared = 7.043, df = 8, p-value = 0.532
```

Comments on Goodness of Fit H0: Our model is adequate H1: Our model is not adequate As we can see shown above, the Hosmer-Lemeshow's Goodness of Fit Test contained a p-value of 0.532, which is clearly above the significance level of 0.10, therefore we cannot reject our null hypothesis (H0), which means the model fits well and is considered to be adequate.

Checking Residual Plots

```
residualdeviance = residuals(stepwiseselectionmale, type = "deviance")
residualpearson = residuals(stepwiseselectionmale, type = "pearson")
standardresidualdeviance = residuals(stepwiseselectionmale, type = "deviance")/sqrt(1 -
standardresidualpearson = residuals(stepwiseselectionmale, type = "pearson")/sqrt(1 -

par(mfrow=c(1,2))
plot(standardresidualdeviance[stepwiseselectionmale$model$LiverPatient==0], col = "red
     ylim = c(-3.5,3.5), ylab = "Standard Deviance Residuals", xlab = "ID")
points(standardresidualdeviance[stepwiseselectionmale$model$LiverPatient==1], col = "b

plot(standardresidualpearson[stepwiseselectionmale$model$LiverPatient==0], col = "red"
     ylim = c(-3.5,3.5), ylab = "Standard Pearson Residuals", xlab = "ID")
points(standardresidualpearson[stepwiseselectionmale$model$LiverPatient==1], col = "bl
```

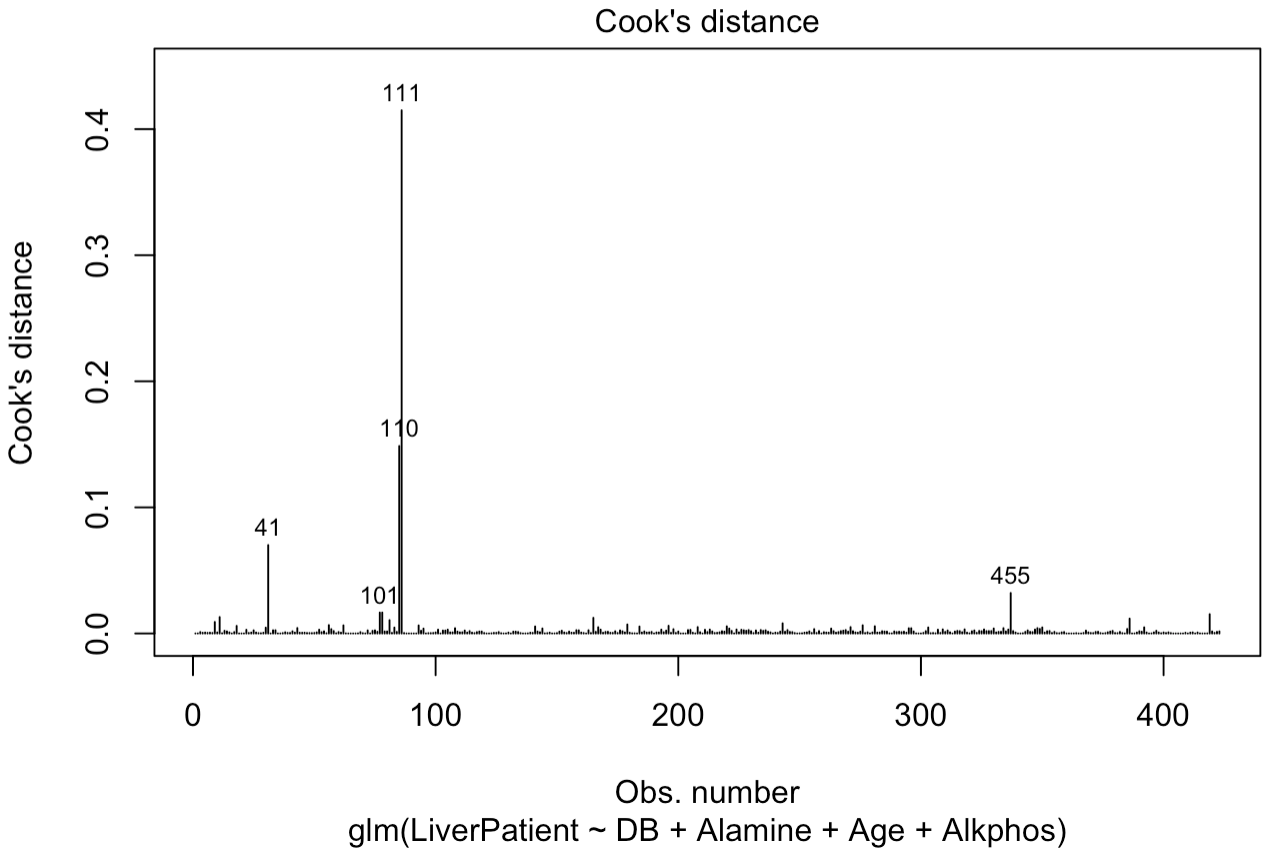




Comments The residual plots follow a similar pattern in both the Standard Deviance Residuals and Standard Pearson Residuals. As we can see above, the red line does not go away from linearity and there is no large outstanding values it does not violate any assumptions. As a result, because there really are not any points with extremely big values, the Bernoulli assumption is valid. In addition, because there is no systematic pattern in the plot, we can conclude that the homoscedasticity assumption is true.

Using Cook's Distance Plot

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



```
influenceddiagnostics2 = which(cooks.distance(stepwiseselectionmale)>0.25)
influenceddiagnostics2
```

```
111
 86
```

Comments As a result from Cook's Distance Plot, we can see there are 2 influential points (observations 111 and 86) made with a Cook's Distance Plot greater than 0.25. We need to eliminate these influential points and refit the Model again.

Refitting Model

```
glmrefitliver2 =  glm(LiverPatient ~ DB + Alamine + Age + Alkphos, data = liverMale[-i
```

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

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). NOTE: stepwise selection with AIC criteria can be performed by default step() function in R. **Final Model**

```
summary(glmrefitliver2)
```

```
Call:
glm(formula = LiverPatient ~ DB + Alamine + Age + Alkphos, family = "binomial",
    data = liverMale[-influenceddiagnostics2, ])

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
```

Comments The final model would be **log(p/1-p) = -1.902754 + 0.573104 * DB + 0.015850 *

Alamine + 0.020418 * Age + 0.003744 * Alkphos

Calculating Odds Ratio

```
round(exp(glmrefitliver2$coefficients),3)
```

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

Interpretation of Odds Ratio As we can see shown above, we make the conclusion that the probability of male being a liver patient with active liver disease can be increased by a factor of The odds of a male liver patient having active liver disease increases by a fact of 1.774 with every unit increase in DB. On the other hand, we can assume all other variables remain constant the odds of liver disease increase by a factor of the following:The odds of liver disease increase by a factor of 1.016 with a one unit increase in Alamine, the odds of liver disease increase by a factor of 1.021 with a one unit increase in Age, and the odds of liver disease increase by a factor of 1.004 with a one unit increase in Alkphos variable. As a result we can make the conclusion, that an adult male with high levels of Direct Bilirubin (DB), Alamine Aminotransferase (Alamine), Alkaline Phosphotase (Alkphos), and Age is more likely to be a liver patient with active liver disease.

d. Comment on how the models for adult females and adult males differ. Use significance level alpha=0.1

Comments As we can see from above, when we compare the models for both females and males we can make the assumption that the females only have two significant predictors (DB and Aspartate) while the males have four significant predictors (DB, Alamine, Age, and Alkphos). While all of these predictors increase the odds of a patient having liver disease it may be easier to start predicting when a male is at risk easier than a female, simply because of the amount of predictors used.

Exercise 3: Use the sleep data set which originates from http://lib.stat.cmu.edu/datasets/sleep. maxlife10 is 0 if the species maximum life span is less than 10 years and 1 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

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

Importing Data

```
sleepdata = read.csv("sleep-1.csv", header = TRUE)
str(sleepdata)
```

```
'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 ...
```

Running Fit Logistic Regression Model

```
glmnullsleep = glm(maxlife10 ~ 1, data = sleepdata, family = "binomial")

glmfullsleep = glm(maxlife10 ~ bodyweight + brainweight + totalsleep + gestationtime +
                      data = sleepdata, family = "binomial")
```

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

Stepwise Selection with AIC Criteria

```
stepwiseselectionsleep = step(glmnullsleep, scope = list(upper=glmfullsleep),
     direction ="both",test ="Chisq", trace = F)
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
summary(stepwiseselectionsleep)
```

```
Call:
glm(formula = maxlife10 ~ brainweight + totalsleep + as.factor(sleepexposureindex) +
    as.factor(predationindex), family = "binomial", data = sleepdata)

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
```

Comments Overall, it is clear that our variable **sleepexposureindex2** is a significant predictor due to the fact that the p-value is lower than the significance level of 0.10. In addition, there would be some discrepancies in model fit, due to the fact there are only 51 observations in the model. Also, because the index variables are considered categorical, the model would struggle to fit as it requires sufficient number of data points for accurate fit for all the categories.

b. What does 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.

Calculating Goodness of Fit Using Hosmer-Lemeshow's Test

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

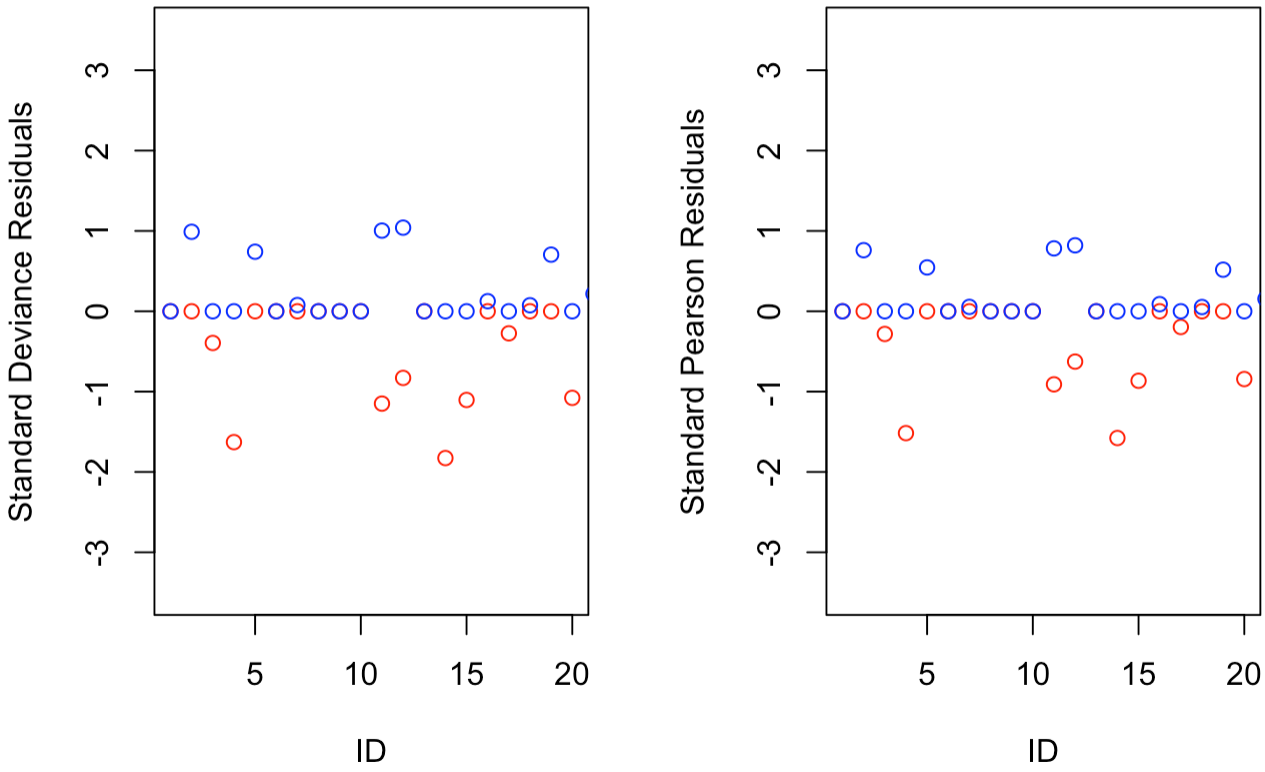
```
Warning in hoslem.test(stepwiseselectionsleep$y,
fitted(stepwiseselectionsleep), : The data did not allow for the requested
number of bins.


    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepwiseselectionsleep$y, fitted(stepwiseselectionsleep)
X-squared = 7.0397, df = 7, p-value = 0.4248
```

Comments on Goodness of Fit H0: Our model is adequate H1: Our model is not adequate As we can see shown above, the Hosmer-Lemeshow's Goodness of Fit Test contained a p-value of 0.4248, which is clearly above the significance level of 0.10, therefore we cannot reject our null hypothesis (H0), which means the model fits well and is considered to be adequate.

Checking Residual Plots

```
residualdeviance2<-residuals(stepwiseselectionsleep, type = "deviance")
```

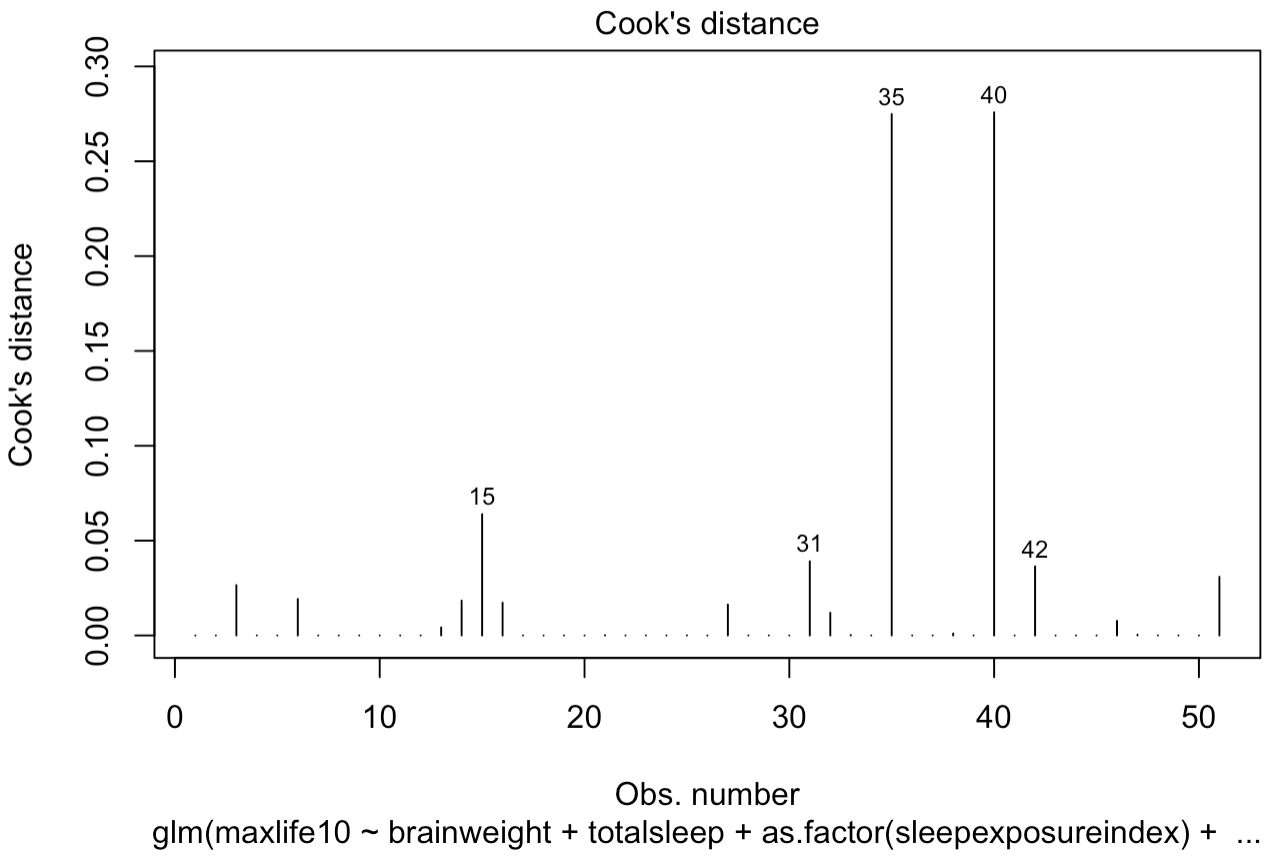
```
residualpearson2<-residuals(stepwiseselectionsleep, type = "pearson")
standardresidualdeviance2<-residuals(stepwiseselectionsleep, type = "deviance")/sqrt(1
standardresidualpearson2 <-residuals(stepwiseselectionsleep, type = "pearson")/sqrt(1

par(mfrow=c(1,2))
plot(standardresidualdeviance2[stepwiseselectionsleep$model$maxlife10==0], col = "red"
points(standardresidualdeviance2[stepwiseselectionsleep$model$maxlife10==1], col = "bl

plot(standardresidualpearson2[stepwiseselectionsleep$model$maxlife10==0], col = "red",
points(standardresidualpearson2[stepwiseselectionsleep$model$maxlife10==1], col = "blu
```



Comments As we can see above, both the standard deviance and standard pearson residual plots, the points are distributed between 2 and -2. Since there is not a pattern in the distribution of the residuals, as a result the model assumptions of residuals having Bernoulli distribution and homoscedasticity is valid. In addition to this, the linearity assumption is not violated since there is not a systematic pattern in the plot.

Using Cook's Distance Plot

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




```
influenceddiagnostics3 = which(cooks.distance(stepwiseselectionsleep)>0.25)
influenceddiagnostics3
```

```
35 40
35 40
```

Comments As a result from Cook's Distance Plot, we can see there are 2 influential points (observations 35 and 40) made with a Cook's Distance Plot greater than 0.25. We need to eliminate these influential points and refit the Model again.

Running Fit Logistic Regression Model

```
glmrefitsleep = glm(maxlife10 ~ brainweight + totalsleep + as.factor(predationindex) +
                      data = sleepdata[-influenceddiagnostics3], family = "binomial")
```

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

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. NOTE: For part (c), interpret the Odds Ratio for all covariates regardless of their significance. **Final Model**

```
summary(glmrefitsleep)
```

```
Call:
glm(formula = maxlife10 ~ brainweight + totalsleep + as.factor(predationindex) +
    as.factor(sleepexposureindex), family = "binomial", data =
sleepdata[-influenceddiagnostics3])

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
```

Calculating Odds Ratio

```
round(exp(glmrefitsleep$coefficients),3)
```

```
                  (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
```

Interpretation of Odds Ratio As we can see shown above, we make the conclusion that the odds(sleepexposureindex = 2)/odds(sleepexposureindex = 1) = exp(4.998e+00) = 1.480500e+02

when the sleep exposure index is 1 when the rest of the other variables are kept constant. With this in mind, 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. As a result, we can make the assumption that animals that sleep in the second-best (Level 2) well-protected den will 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 a)-c) by treating two index variables as continuous variables.

a. First find and specify the best set of predictors via stepwise selection with AIC criteria. **Running Fit Logistic Regression Model**

```
glmnullsleep2 = glm(maxlife10 ~ 1, data = sleepdata, family = "binomial")

glmfullsleep2 = glm(maxlife10 ~ bodyweight + brainweight + totalsleep + gestationtime
```

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

Stepwise Selection with AIC Criteria

```
stepwiseselectionsleep2 = step(glmnullsleep2, scope = list(upper=glmfullsleep2),
     direction ="both",test ="Chisq", trace = F)
```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```r
summary(stepwiseselectionsleep2)
```

```
Call:
glm(formula = maxlife10 ~ brainweight + totalsleep + sleepexposureindex +
    predationindex, family = "binomial", data = sleepdata)

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
```

```
brainweight         0.00018    0.03544   1.098   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
```

Comments Overall, it is clear that our variables **brainweight, totalsleep, sleepexposureindex, and predationindex** are significant predictors due to the fact that their p-values are lower than the significance level of 0.10. In addition to this, we will consider both our variables **sleep exposure index and predation index** as continuous variable that cater the data issue discussed in Exercise 3.

b. What does 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. **Calculating Goodness of Fit Using Hosmer-Lemeshow's Test**

```
hoslem.test(stepwiseselectionsleep2$y, fitted(stepwiseselectionsleep2), g=10)
```

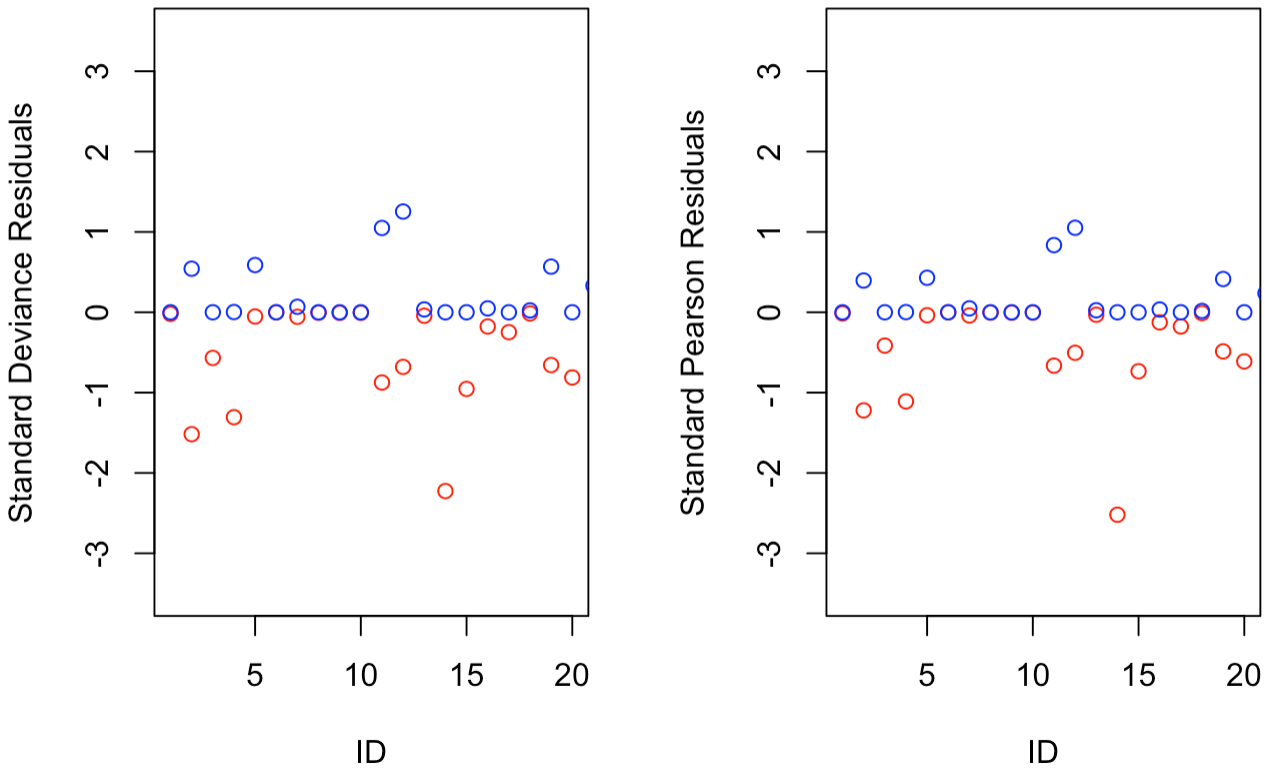
```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepwiseselectionsleep2$y, fitted(stepwiseselectionsleep2)
X-squared = 1.4406, df = 8, p-value = 0.9937
```

Comments on Goodness of Fit H0: Our model is adequate H1: Our model is not adequate As we can see shown above, the Hosmer-Lemeshow's Goodness of Fit Test contained a p-value of 0.9937, which is clearly above the significance level of 0.10, therefore we cannot reject our null hypothesis (H0), which means the model fits well and is considered to be adequate. Our variables, **brainweight, totalsleep, sleepexposureindex, and predationindex** these predictors have p-values of 0.0895, 0.0865, 0.0252, and 0.0265 are clearly below the significance level of 0.10; as a result there is a significant relationship between **brainweight, totalsleep, sleepexposureindex, and predationindex** and this means whether an animal species' maximum lifespan will be at least 10 years.

Checking Residual Plots

```
residudaldeviance3<-residuals(stepwiseselectionsleep2, type = "deviance")
residualpearson3<-residuals(stepwiseselectionsleep2, type = "pearson")
standardresidualdeviance3<-residuals(stepwiseselectionsleep2, type = "deviance")/sqrt(
```

```
standardresidualpearson3 <-residuals(stepwiseselectionsleep2, type = "pearson")/sqrt(1

par(mfrow=c(1,2))
plot(standardresidualdeviance3[stepwiseselectionsleep2$model$maxlife10==0], col = "red
points(standardresidualdeviance3[stepwiseselectionsleep2$model$maxlife10==1], col = "b

plot(standardresidualpearson3[stepwiseselectionsleep2$model$maxlife10==0], col = "red"
points(standardresidualpearson3[stepwiseselectionsleep2$model$maxlife10==1], col = "bl
```

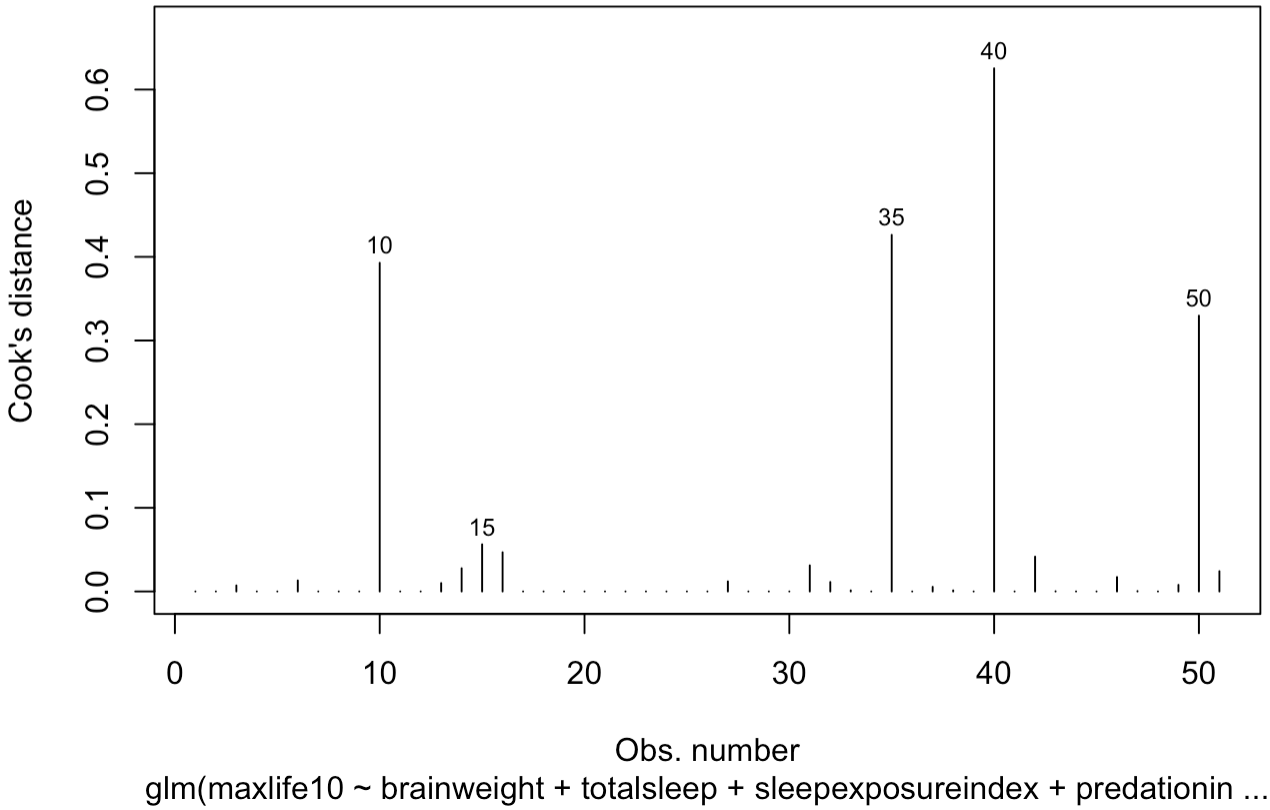


Comments As we can see above, both the standard deviance and standard pearson residual plots, the points are distributed between 2 and -2. Since there is not a pattern in the distribution of the residuals, as a result the model assumptions of residuals having Bernoulli distribution and homoscedasticity is valid. In addition to this, the linearity assumption is not violated since there is not a systematic pattern in the plot.

Using Cook's Distance Plot

```
plot(stepwiseselectionsleep2, which = 4, id.n = 5)
```

Cook's distance




```
influenceddiagnostics4 = which(cooks.distance(stepwiseselectionsleep2)>0.25)
influenceddiagnostics4
```

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

Comments As a result from Cook's Distance Plot, we can see there are 4 influential points (observations 10, 35, 40, 50) made with a Cook's Distance Plot greater than 0.25. We need to eliminate these influential points and refit the Model again.

Running Fit Logistic Regression Model

```
glmrefitsleep2 = glm(maxlife10 ~ brainweight + totalsleep + as.factor(predationindex)
                     data = sleepdata[-influenceddiagnostics4], family = "binomial")
```

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

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. NOTE: For part (c), interpret the Odds Ratio for all covariates regardless of their significance. **Final Model**

```
summary(glmrefitsleep2)
```

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

```
glm(formula = maxlife10 ~ brainweight + totalsleep + as.factor(predationindex) +
    as.factor(sleepexposureindex), family = "binomial", data =
sleepdata[-influenceddiagnostics4])

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
```

Calculating Odds Ratio

```
round(exp(glmrefitsleep2$coefficients),3)
```

```
                  (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
```

Interpretation of Odds Ratio As we can see shown above, we make the conclusion that the odds that a species' maximum lifespan will be at least 10 years increase by a factor of exp(5.101e-02) = 1.052000e+00 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(4.230e-01) = 1.527000e+00 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(-2.535e+00) = 7.900000e-02 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(-2.535e+00) = 1.480500e+02 with a one unit increase in sleepexposureindex when totalsleep, brainweight, and predationindex are held constant.

Therefore as a result 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. In other words, 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.