**Homework 5**

**Exercise 1:**

1. **For females in the data set, determine the best set of predictors for a logistic regression model predicting whether a female is a liver patient, and comment on any unduly influential points. If any extremely unduly influential points exist, remove them and perform selection again before choosing a final model. The cut-off for Cbar measure is set as 0.5.**

For females in the dataset we run the query builder to filter females from gender variable. There are 135 observations for female gender

We run logistic regression with stepwise model with 135 observations for female gender with 9 predictors and 1 binary/dichotomous response variable(LiverPatient). Our event of interest is in checking for LiverPatient = 1

We observe the global null hypothesis to check the coefficients of regression for all predictors. The Likelihood ratio has a p-value 0.0002 which is significant at p < 0.05 This means that at least one β is not zero and the logistic regression model is useful. Since there is only one significant predictor from the model we can conclude that the β of Aspartate is not 0

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 13.7678 | 1 | 0.0002 |
| **Score** | 7.3489 | 1 | 0.0067 |
| **Wald** | 5.9965 | 1 | 0.0143 |

The below table shows that regression coefficient of Aspartate has p-value 0.0143 significant at p < 0.05, we reject the null and conclude that **regression coefficient is significantly different from 0.**

Hence only **predictor Aspartate is making a significant contribution** to predict the odds of a female adult being a liver patient

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.1422 | 0.3056 | 0.2165 | 0.6417 |
| **Aspartate** | 1 | 0.0169 | 0.00689 | 5.9965 | 0.0143 |

**We do not find** **any extremely unduly influential points with cbar > 0.5**



1. **If any points are still too influential in your final model, remove them and refit. Comment on the significance of parameter estimates, what Hosmer-Lemeshow’s test tells us about goodness of fit, and point out any remaining issues with diagnostics.**

**We do not find** **any extremely unduly influential points with cbar > 0.5**

The global null hypothesis shows that the likelihood ratio is significant at p<0.05 hence we reject null. At least one β is not 0

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 13.7678 | 1 | 0.0002 |
| **Score** | 7.3489 | 1 | 0.0067 |
| **Wald** | 5.9965 | 1 | 0.0143 |

The below table shows that regression coefficient of Aspartate has p-value 0.0143 significant at p < 0.05, we reject the null and conclude that **regression coefficient is significantly different from 0.**

Hence the **predictor Aspartate is making a significant contribution** to predict the odds of a female adult being a liver patient

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.1422 | 0.3056 | 0.2165 | 0.6417 |
| **Aspartate** | 1 | 0.0169 | 0.00689 | 5.9965 | 0.0143 |

The hosmer-Lemeshow’s goodness-of-fit test gives a p-value 0.7828 greater than 0.05 hence we fail to reject the null and conclude that the **model is adequate**.

| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| --- | --- | --- |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 4.7609 | 8 | 0.7828 |

The influence diagnostics plot for pearson and deviance residual shows majority of the observations between -2 and 2 and the cbar diagnostics has no influential points greater than the cutoff 0.5

Hence they are no issues with the diagnostics.





1. **Comment on the significance of odds ratios and interpret what the model tells us about relationships between the predictors and the odds of an adult female being a liver patient.**

From the below table , odds ratio Aspartate is significant as its confidence intervals does not contain 1

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **Aspartate** | 1.017 | 1.003 | 1.031 |

From the above table ,

Odds ratio for Aspartate is 1.017 that is exp(0.0169). The odds of a female adult being a liver patient changes by factor of 1.017 for every 1 unit increase in Aspartate. Aspartate has a positive relationship with the odds of an adult female being a liver patientsince its β 0.0169 is greater than 0

As Aspartate increases by 1 unit, the odds of an adult female being a liver patient increases by a factor of 1.017

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.1422 | 0.3056 | 0.2165 | 0.6417 |
| **Aspartate** | 1 | 0.0169 | 0.00689 | 5.9965 | 0.0143 |

**Exercise 2:**

**Repeat exercise 1 for males. In addition to the previous questions, also comment on how the models for adult females and adult males differ. Again, the cut-off for Cbar measure is set as 0.5.**

For males in the dataset we run the query builder to filter males from gender variable. There are 423 observations for male gender

**We find** **one extremely unduly influential point with cbar\_LiverPatient > 0.5**



After removing the influential point, we run logistic regression with full fitted model for **422** observations of male gender with 4 significant predictors and 1 binary/dichotomous response variable (LiverPatient). Our interest is in checking for LiverPatient = 1

**We do not find any more unduly influential point with cbar > 0.5**

Age, DB, Alamine, AGRatio are the best set of predictors for our logistic regression model predicting whether a male is a a liver patient

The Likelihood ratio has a p-value < .0001 which is significant at p < 0.05 This means that at least one β is not zero and the logistic regression model is useful.

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 88.7555 | 4 | <.0001 |
| **Score** | 42.3988 | 4 | <.0001 |
| **Wald** | 37.2145 | 4 | <.0001 |

The below table shows that regression coefficient of Age, DB, ALamine, AGRatio has p-value 0.0163, 0.0035, 0.0002, 0.0351 respectively, significant at p < 0.05, we reject the null and conclude that regression coefficients are significantly different from 0.

Hence the corresponding **predictors Age, DB, ALamine, AGRatio make a significant contribution to predict the odds of an adult male being a liver patient.**

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.3894 | 0.6684 | 0.3395 | 0.5601 |
| **Age** | 1 | 0.0199 | 0.00831 | 5.7688 | 0.0163 |
| **DB** | 1 | 0.5617 | 0.1925 | 8.5149 | 0.0035 |
| **Alamine** | 1 | 0.0223 | 0.00592 | 14.2321 | 0.0002 |
| **AGRatio** | 1 | -0.8912 | 0.4229 | 4.4406 | 0.0351 |

From the below table , the odds ratio of all 4 predictors is significant as their confidence intervals does not contain 1

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **Age** | 1.020 | 1.004 | 1.037 |
| **DB** | 1.754 | 1.203 | 2.558 |
| **Alamine** | 1.023 | 1.011 | 1.035 |
| **AGRatio** | 0.410 | 0.179 | 0.940 |

From the below table ,

Odds ratio for age is 1.020 that is exp(0.0199). The odds of a male adult being a liver patient changes by factor of 1.020 for every 1 unit increase in age. Age has a positive relationship with odds of an adult male being a liver patient since its β 0.0199 is greater than 0

As age increases by 1 unit the odds of a male adult being a liver patient increases by a factor of 1.020.

Odds ratio for DB is 1.754 that is exp(0.5617). The odds of a male adult being a liver patient changes by factor of 1.754 for every 1 unit increase in DB. DB has a positive relationship with with odds of an adult male being a liver patient since its β 0.5617 is greater than 0

As DB increases by 1 unit the odds of a male adult being a liver patient increases by a factor of 1.754.

Odds ratio for Alamine is 1.023 that is exp(0.0223). The odds of a male adult being a liver patient changes by factor of 1.023 for every 1 unit increase in alamine. Alamine has a positive relationship with odds of an adult male being a liver patient since its β 0.0223 is greater than 0

As alamine increases by 1 unit the odds of a male adult being a liver patient increases by a factor of 1.023

Odds ratio for AGRatio is 0.410 that is exp(-0.8912). The odds of a male adult being a liver patient changes by factor of 0.410 for every 1 unit increase in AGRatio. AGRatio has a negative relationship with odds of an adult male being a liver patient since its β -0.8912 is less than 0

As AGRatio increases by 1 unit the odds of a male adult being a liver patient decreases by a factor of 0.410

The hosmer-Lemeshow’s goodness-of-fit test gives a p-value 0.5696 greater than 0.05 hence we fail to reject the null and conclude that the model is adequate

| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| --- | --- | --- |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 6.6975 | 8 | 0.5696 |

The influence diagnostics plot for Pearson and deviance residual shows majority of the observations between -2 and 2 and the cbar diagnostics has no influential points greater than the cutoff 0.5

Hence they are no issues with the diagnostics for the final model





**Difference between model for adult female and adult male** :

Only the predictor Asparate makes a significant contribution in predicting the odds of a female adult being a liver patient whereas Asparate is insignificant in predicting the odds of male adult being a liver patient instead Age, DB, Alamine, AGRatio are significant in prediciting the odds of a male adult being a liver patient. We do not find any extremely unduly influential points with cbar > 0.5 for female adults but we find 1 influential point with cbar > 0.5 for male adults.

**Exercise 3:**

1. **First determine and specify the best set of predictors through stepwise selection with criteria .05. Comment on any unduly influential points on the final model. Use cut-off as 1.5. If any extremely unduly influential points exist, remove them for the final model.**

Maxlife 10 is the binary response variable. We are interested in event that species’ max lifespan will be at least 10 years (maxlife 10 = 0)

Bodyweight, brianweight, totalsleep, gestationtime are the quantitative/continuous predictors

Predationindex, sleepexposureindex, overalldangerindex are the categorical predictors

From stepwise selection, we have only one significant predictor, gestationtime with p-value 0.0035 siginificant at p < 0.05 .

| **Type 3 Analysis of Effects** | | | |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **gestationtime** | 1 | 8.5131 | 0.0035 |

We run full fitted model with only significant predictor gestationtime The Likelihood ratio has a p-value < .0001 which is significant at p < 0.05 This means that β of gestationtime is not zero and the logistic regression model is useful.

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 15.3506 | 1 | <.0001 |
| **Score** | 10.7615 | 1 | 0.0010 |
| **Wald** | 8.2827 | 1 | 0.0040 |

**There are no extremely unduly influential points above cbar\_maxlife10 > 1.5**



1. **Comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any remaining issues with diagnostics.**

The global null hypothesis shows that the likelihood ratio is significant at p<0.05 hence we reject null. At least β of gestationtime is not 0

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 15.3506 | 1 | <.0001 |
| **Score** | 10.7615 | 1 | 0.0010 |
| **Wald** | 8.2827 | 1 | 0.0040 |

The below table shows that regression coefficient of gestationtime has p-value 0.0040 significant at p < 0.05, we reject the null and conclude that regression coefficient is significantly different from 0.

Hence the predictor gestationtime is making a significant contribution to predict the odds of a species' maximum lifespan being at least 10 years

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.8759 | 0.4632 | 3.5760 | 0.0586 |
| **gestationtime** | 1 | 0.0119 | 0.00414 | 8.2827 | 0.0040 |

From Hosmer-Lemeshow's test, we get a p-value 0.4072 greater than 0.05, we fail to reject null and conclude that the model is adequate.

| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| --- | --- | --- |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 8.2739 | 8 | 0.4072 |

**Diagnostics plots**: Pearson and deviance residual show most of the observations between -2 to 2, cbar measure shows no points greater than 1.5 hence no extremely unduly influential points. No issues with diagnostics plots





1. **Comment on the significance of odds ratios and interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years.**

The odds ratio for gestationtime is significant as its confidence interval does not include 1

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **gestationtime** | 1.012 | 1.004 | 1.020 |

Logistic regression line: Log(p/1-p) = -0.8759+ 0.0119 \* gestationtime

The odds of a species' maximum lifespan being at least 10 years changes by a factor of 1.012 with one unit increase in gestationtime.

gestationtime has a positive relationship with the odds of a species' maximum lifespan being at least 10 years since its β 0.0119 is greater than 0

As gestationtime increases by 1 unit, the odds of a species' maximum lifespan being at least 10 years increases by a factor of 1.012

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.8759 | 0.4632 | 3.5760 | 0.0586 |
| **gestationtime** | 1 | 0.0119 | 0.00414 | 8.2827 | 0.0040 |

**Exercise 4:**

1. **Determine and specify the best set of predictors through stepwise selection with criteria .05. Comment on any unduly influential points on the final model. Use cut-off as 1.5. If any extremely unduly influential points exist, remove them for the final model.**

From stepwise selection we find 2 significant predictors, sleepexposureindex and predationindex with p-values <.0001 significant at p < 0.05

| **Summary of Stepwise Selection** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Effect** | | **DF** | **Number In** | **Score Chi-Square** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Entered** | **Removed** |
| **1** | **sleepexposureindex** |  | 1 | 1 | 18.5314 |  | <.0001 |
| **2** | **predationindex** |  | 1 | 2 | 18.3581 |  | <.0001 |

We remove is **1 influential point** with cbar\_maxlife10 > 1.5 and have 61 observations in the final model



1. **Comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any remaining issues with diagnostics.**

We run full fitted model for the both predictors

The global null hypothesis shows that the likelihood ratio is significant at p<0.05 hence we reject null. At least one β is not 0

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 50.9260 | 2 | <.0001 |
| **Score** | 32.5994 | 2 | <.0001 |
| **Wald** | 12.3609 | 2 | 0.0021 |

The below table shows that both regression coefficients of predationindex and sleepexposureindex with p-values 0.0012 and 0.0004 are significant at p < 0.05 we reject null and conclude that both the regression coefficient are significantly different from 0.

Hence both the corresponding predictors make a significant contribution to predict the odds of a species' maximum lifespan being at least 10 years.

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.0256 | 0.8524 | 0.0009 | 0.9760 |
| **predationindex** | 1 | -2.9882 | 0.9204 | 10.5402 | 0.0012 |
| **sleepexposureindex** | 1 | 3.7457 | 1.0654 | 12.3605 | 0.0004 |

From Hosmer-Lemeshow's test, we get a p-value 0.9924 greater than 0.05, we fail to reject null and conclude that the model is adequate.

| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| --- | --- | --- |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 1.1307 | 7 | 0.9924 |

**Diagnostics plots**: Pearson and deviance residual show most of the observations between -2 to 2, cbar measure shows no points greater than 1.5 hence no extremely unduly influential points in the final model. No issues with diagnostics plots





1. **Comment on the significance of odds ratios and interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years**.

The odds ratio for **predationindex** and **sleepexposureindex** are significant as their confidence intervals do not include 1

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **predationindex** | 0.050 | 0.008 | 0.306 |
| **sleepexposureindex** | 42.338 | 5.246 | 341.662 |

Logistic regression line: Log(p/1-p) = -0.0256 - 2.9882\* predationindex + 3.7457 \* sleepexposureindex

The odds of a species' maximum lifespan being at least 10 years changes by a factor of 0.050 with one unit increase in predationindex. predationindex has a negative relationship with the odds of a species' maximum lifespan being at least 10 years since its β -2.9882 is less than 0

As predationindex increases by 1 unit, the odds of a species' maximum lifespan being at least 10 years decreases by a factor of 0.050

The odds of a species' maximum lifespan being at least 10 years changes by a factor of 42.338 with one unit increase in sleepexposureindex. sleepexposureindex has a positive relationship with the odds of a species' maximum lifespan being at least 10 years since its β 3.7457 is greater than 0

As sleepexposureindex increases by 1 unit, the odds of a species' maximum lifespan being at least 10 years increases by a factor of 42.338

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -0.0256 | 0.8524 | 0.0009 | 0.9760 |
| **predationindex** | 1 | -2.9882 | 0.9204 | 10.5402 | 0.0012 |
| **sleepexposureindex** | 1 | 3.7457 | 1.0654 | 12.3605 | 0.0004 |