Applied Logistic Regression - Exercise Week 5

Yago Durán Cid

13/06/2015

WEEK 5

Exercise 1:

Use the hyponatremia.dta dataset to complete the following
a. Assess the association between hyponatremia (dichotomous variable nas135) and sex (variable female) by making a 2 by 2 table. Calculate the odds ratio of hyponatremia of a female compared to a male. Compute the 95% confidence interval for this odds ratio. Interpret the findings.

```
##         nas135   0   1
## female
## 0              297  25
## 1              129  37
```

The Odds ratio is 3.4074419 and the log of odd ratio is 1.2259618
The Variance is 0.078146 and the Std. Deviation is 0.279546

To obtain the 95% confidence interval for the log of Odd Ratio we apply $log(Odd\,Ratio) \pm 1.96 * Std.\,Deviation$

Thus, the upper bound is 1.773872 and the lower bound is 0.6780517

Exponentiating we get the limits of the Odd Ratio. Upper bound is 5.8936293 and the lower bound is 1.9700357

The odds of a female experiencing hyponatremia is 3.4 times greater than that of a male. The 95% Confidence interval for the odds ratio is (1.97, 5.89). Upon repeated sampling, 95% of confidence intervals constructed this way would cover the true population odds ratio.

b. Perform a logistic regression analysis with R using nas135 as dependent variable and female as the only independent variable. Use the Likelihood Ratio test to assess the significance of the model. Is the model with female a better model than the naïve model?

```
##
## Call:
## glm(formula = nas135 ~ female, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7102  -0.7102  -0.4020  -0.4020   2.2608
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.4749     0.2082 -11.884  < 2e-16 ***
## female        1.2260     0.2795   4.386 1.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 371.60  on 487  degrees of freedom
## Residual deviance: 351.93  on 486  degrees of freedom
## AIC: 355.93
##
## Number of Fisher Scoring iterations: 5
```

The p-value of the likelihood test comparing the model with female and the one without is 9.2039008×10^{-6} The model with female is significantly better than the naïve model.

c. What is the naïve model? What is the probability of hyponatremia that this model predict?

The naïve model (excluding any independent variable) can be seen in the table below

```
##    0   1
##
##  426  62
```

The naïve model predicts a $\frac{62}{426+62} = 12.704918\%$ probability of hyponatremia for every subject in the study.

d. Run a logistic regression analyses with no independent variables. Transform the coefficient obtained from this model into a probability.

```
##
## Call:
## glm(formula = nas135 ~ 1, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5213  -0.5213  -0.5213  -0.5213   2.0313
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9273     0.1359  -14.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 371.6  on 487  degrees of freedom
## Residual deviance: 371.6  on 487  degrees of freedom
## AIC: 373.6
##
## Number of Fisher Scoring iterations: 4
```

The odds is $Odds = \frac{Prob_{x=1}}{1-Prob_{x=1}}$. Manipulating that equation we can get that $Prob_{x=1} = \frac{Odds}{1+Odds}$

The logit is the log(Odds), exponentiating we can get the probability from equation above $Prob_{x=1} = \frac{e^{logit}}{1+e^{logit}}$

Using the value estimated in our model the probaility is 0.1270492

e. Using the model with female as independent variable, compute the estimated probability of hyponatremia per males and females. Write down the equation for the logit.

The model including female as independent variable is

```
##
## Call:
## glm(formula = nas135 ~ female, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7102  -0.7102  -0.4020  -0.4020   2.2608
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.4749     0.2082 -11.884  < 2e-16 ***
## female        1.2260     0.2795   4.386 1.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 371.60  on 487  degrees of freedom
## Residual deviance: 351.93  on 486  degrees of freedom
## AIC: 355.93
##
## Number of Fisher Scoring iterations: 5
```

The logit is $g(female) = \beta_0 + \beta_1 female$

Using the formula above, the probability of hyponatremia for a female is $Prob_{nas135=1} = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} = 0.2228916$

the probability of hyponatremia for a male is $Prob_{nas135=1} = \frac{e^{\beta_0}}{1+e^{\beta_0}} = 0.0776398$

f. Use the Wald test to assess the significance of the coefficient for female.

The Wald test is given in the summary of the model above.

g. Fit a model with runtime as the only independent variable. Assess the significance of the model.

```
##
## Call:
## glm(formula = nas135 ~ runtime, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1724  -0.5234  -0.4263  -0.3458   2.5182
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.592594   0.771282  -7.251 4.14e-13 ***
## runtime      0.015502   0.003091   5.015 5.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 360.90  on 476  degrees of freedom
## Residual deviance: 335.54  on 475  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 339.54
##
## Number of Fisher Scoring iterations: 5
```

The p-value of the likelihood test comparing the model with runtime and the one without it is 4.7770591×10^{-7} The model with runtime is significantly better than the naïve model.

h. Calculate the probability of hyponatremia of a runner who takes 4 hours (240 minutes) to complete the marathon.

The value of the logit for a runner taking 4 hours is -1.872131 and the probablity following the equation described above is 0.1332953

i. Fit a model with female and runtime as independent variables. Assess the significance of the model. Which null hypothesis is tested?

```
##
## Call:
## glm(formula = nas135 ~ runtime, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1724  -0.5234  -0.4263  -0.3458   2.5182
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.592594   0.771282  -7.251 4.14e-13 ***
## runtime      0.015502   0.003091   5.015 5.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 360.90  on 476  degrees of freedom
## Residual deviance: 335.54  on 475  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 339.54
##
## Number of Fisher Scoring iterations: 5
```

The p-value of the likelihood test comparing the model with female and runtime and the one without them is 3.1255477×10^{-6} The model with female is significantly better than the naïve model.