Applied Logistic Regression - Exercise Week 3

Yago Durán Cid

30/05/2015

WEEK 3

Exercise 1: Use the Myopia Study (MYOPIA.csv)

a. Using the results from Week 2, Exercise 1, compute 95 percent confidence intervals for the slope coefficient SPHEQ. Write a sentence interpreting this confidence.

We keep the basic model we used in homework 2. This is $\pi(x) = E(y|x) = \frac{e^{(\beta_0+\beta_1 SPHEQ)}}{1+e^{(\beta_0+\beta_1 SPHEQ)}}$ which gives foolowing results:

```
##
## Call:
## glm(formula = MYOPIC ~ SPHEQ, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6435  -0.4533  -0.2681  -0.1029   3.1602
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05397    0.20675   0.261    0.794
## SPHEQ       -3.83310    0.41837  -9.162   <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: 480.08  on 617  degrees of freedom
## Residual deviance: 337.34  on 616  degrees of freedom
## AIC: 341.34
##
## Number of Fisher Scoring iterations: 6
```

As we know, the confidence interval can be estimated as $\beta_j \pm z_{\frac{1-\alpha}{2}}\hat{\sigma_{\beta_j}}$

Thus, the higher bound of the confidence interval is $-3.83310 + 1.96\sqrt{0.17503316}$=-3.0130931
And the lower bound is $-3.83310 - 1.96\sqrt{0.17503316}$=-4.6531021

Given the confidence interval estimated above we can say that, with 95% confidence, the true value of β_{SPHEQ} is between -3.0130931 and -4.6531021 and, therefore, is different than zero (i.e.: the value of SPHEQ impacts the probability of MYOPIA being 1 or 0)

b. Use R to obtain the estimated covariance matrix. Compute the logit and estimated logistic probability for a subject with SPHEQ = 2. Evaluate the endpoints of the 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence interpreting the estimated probability and its confidence interval.

The variance-covariance matrix of the model is:

```
##             (Intercept)       SPHEQ
## (Intercept)  0.04274486 -0.06337768
## SPHEQ       -0.06337768  0.17503316
```

Based on the estimated model, we can obtain the logit for SPHEQ=2

```
##    estimate Std.Error
## 1 -7.612222 0.6995475
```

Substituting in the probability fucntion $Prob(SPHEQ = 2) = \frac{e^{0.0004941278 \pm 1.96 * 0.0003454951}}{1+e^{0.0004941278 \pm 1.96 * 0.0003454951}}$

Average probability of MYOPIA=1 given that SPHEQ=2 is 0.0494128%
At 95% confidence, the probability of MYOPIA=1 given that SPHEQ=2 is between 0.1943908% and 0.0125468%

Exercise 2: Use the ICU study (icu.csv) a. Using the results from Week 1, Exercise 2, part (d), compute 95 percent confidence intervals for the slope and constant term. Write a sentence interpreting the confidence interval for the slope.

As in previous section we show the results of model estimated.

```
## 
## Call:
## glm(formula = STA ~ AGE, family = "binomial", data = icu)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.9536  -0.7391  -0.6145  -0.3905   2.2854  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.05851    0.69608  -4.394 1.11e-05 ***
## AGE          0.02754    0.01056   2.607  0.00913 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 192.31  on 198  degrees of freedom
## AIC: 196.31
## 
## Number of Fisher Scoring iterations: 4
```

The confidence interval at 95% probablity for both the intercept and the variable AGE is:

```
##                   2.5 %      97.5 %
## (Intercept) -4.42280739 -1.69421908
## AGE          0.00683723  0.04824799
```

b. Obtain the estimated covariance matrix for the model fit from Week 1,

```
##              (Intercept)           AGE
## (Intercept)  0.484529087 -0.0071029945
## AGE         -0.007102994  0.0001116015
```

d. Compute the logit and estimated logistic probability for a 60-year old subject. Compute a 95 percent confidence intervals for the logit and estimated logistic probability. Write a sentence or two interpreting the estimated probability and its confidence interval.

```
##    estimate Std.Error
## 1 -1.405957 0.1842154
```

Given that the value of the logit is -1.4059567 with standar error 0.1842154 we can estimate the probability as we did in first section.

Average probability of STA=1 given that AGE=60 is 19.6872578%

At 95% confidence, the probability of STA=1 given that AGE=60 is between 26.0206709% and 14.5913451%

Exercise 3:
Use the ICU study (icu.csv)
Use the ICU data and consider the multiple logistic regression model of vital status, STA, on age (AGE), cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and race (RACE).
a. The variable RACE is coded at three levels. Prepare a table showing the coding of the two design variables necessary for including this variable in a logistic regression model.

In the original dataset, RACE variable is included as integer, which is incorrect

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.175   1.000   3.000
```

We have to convert RACE variable from integer to factor

```
##   1   2   3
## 175  15  10
```

b. Write down the equation for the logistic regression model of STA on AGE, CAN, CPR, INF, and RACE. Write down the equation for the logit transformation of this logistic regression model. How many parameters does this model contain?

$\pi(x) = \frac{e^{\beta_0+\beta_1 AGE+\beta_2 CAN+\beta_3 CPR+\beta_4 INF+\beta_5 RACE2+\beta_6 RACE3}}{1+e^{\beta_0+\beta_1 AGE+\beta_2 CAN+\beta_3 CPR+\beta_4 INF+\beta_5 RACE2+\beta_6 RACE3}}$
The logit is

$g(x) = \beta_0 + \beta_1 AGE + \beta_2 CAN + \beta_3 CPR + \beta_4 INF + \beta_5 RACE2 + \beta_6 RACE3$

We have, thus, seven parameters to estimate.

c. Write down an expression for the likelihood and log likelihood for the logistic regression model in part (b). How many likelihood equations are there? Write down an expression for a typical likelihood equation for this problem.

The likelihood would be as follows:
$\ell(\beta) = \Pi_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i)^{1-y_i}$

Where $y_i = 1$ if STA=1 and $y_i = 0$ otherwise.

Unsing logarithms we get thee log likelihood:
$log(\ell(\beta)) = \sum_{i=1}^{n}((y_i log(\pi(x_i)) + (1-y_i) log(1-\pi(x_i)))$

d. Using a logistic regression package, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part (b). Using these estimates write down the equation for the fitted values, that is, the estimated logistic probabilities.

The estimated parameters are as follows:

```
##
## Call:
## glm(formula = STA ~ AGE + CAN + CPR + INF + RACE, family = "binomial",
##     data = icu2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3703  -0.6823  -0.5421  -0.3082   2.5124
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.51152    0.81443  -4.312 1.62e-05 ***
## AGE          0.02712    0.01159   2.340  0.01926 *
## CAN          0.24451    0.61681   0.396  0.69180
## CPR          1.64650    0.62341   2.641  0.00826 **
## INF          0.68067    0.38042   1.789  0.07357 .
## RACE2       -0.95708    1.08445  -0.883  0.37748
## RACE3        0.25975    0.87127   0.298  0.76561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 179.30  on 193  degrees of freedom
## AIC: 193.3
##
## Number of Fisher Scoring iterations: 5
```

Which translates in
$\pi(x_i) = \frac{e^{-3.51152+0.02712AGE_i+0.24451CAN_i+1.64650CPR_i+0.68067INF_i-0.95708RACE2_i+0.25975RACE3_i}}{1+e^{-3.51152+0.02712AGE_i+0.24451CAN_i+1.64650CPR_i+0.68067INF_i-0.95708RACE2_i+0.25975RACE3_i}}$

e. Using the results of the output from the logistic regression package used in part (d), assess the significance of the slope coefficients for the variables in the model using the likelihood ratio test. What assumptions are needed for the p-values computed for this test to be valid? What is the value of the deviance for the fitted model?

In the likelihood test the null hypotheis is: $H_0 : \forall \beta_i = 0$

$G = Deviance_{model\,without\,variables} - Deviance_{model\,with\,variables} \rightsquigarrow \chi^2_{df}$

In our case, G=200.1609694-179.3007274=20.860242

The p-value for the model taking a $\chi^2_{df=6} = 0.0019437$ Being the p-value<0.05 we reject the null hypothesis.

f. Use the Wald statistics to obtain an approximation to the significance of the individual slope coefficients for the variables in the model. Fit a reduced model that eliminates those variables with nonsignificant Wald statistics. Assess the joint (conditional) significance of the variables excluded from the model. Present the results of fitting the reduced model in a table.

Reviewing the significance obtained fr parameters in section d we can state that variables CAN and RACE are not significant. AGE and CPR are significant at 95% confidence while INF is open to debate. Below we show details for the model inlcluding AGE, CPR and INF variables.

```
##
## Call:
## glm(formula = STA ~ AGE + CPR + INF, family = "binomial", data = icu2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3633  -0.6810  -0.5524  -0.3091   2.4868
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.57604    0.77306  -4.626 3.73e-06 ***
## AGE          0.02792    0.01136   2.458  0.01397 *
## CPR          1.63066    0.61553   2.649  0.00807 **
## INF          0.69708    0.37750   1.847  0.06481 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 180.51  on 196  degrees of freedom
## AIC: 188.51
##
## Number of Fisher Scoring iterations: 5
```

In order to furtherr assess the significance of INF (which p-value remains higher than 0.05 but lower than 0.1) we can perform a likelihood test comparing the model with no INF to the model including INF.

G=200.1609694-180.5134271=3.4392532

The p-value for INF taking a $\chi^2_{df=1}$ = 0.0636645 Being the p-value>0.05 we can discard INF variable at 95% confidence.

g. Using the results from part (f), compute 95 percent confidence intervals for all coefficients in the model. Write a sentence interpreting the confidence intervals for the non-constant covariates.

The 95% confidence interval for the parameters estimated in our last and shortest models are:

```
##                    2.5 %      97.5 %
## (Intercept) -4.813107596 -1.89080341
## AGE          0.007755898  0.05145892
## CPR          0.593811578  2.97437261
```