Neil Yetz

04/19/2016

ERHS 642: Applied Logistic Regression

Dr. Bachand

**ERHS 642 Logistic Regression Spring 2016**

**Homework Assignment 8**

Consider the **ICU\_altered** data set.

1.

1. Present a table containing coefficients, standard errors, Wald Chi-Square values and p-values for your final model from homework assignment 6

Table 1.1: Table presenting coefficients, standard errors, Wald Chi-Square values, and p-values for Final model from homework assignment 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient | Standard Error | Wald Chi-Square | P-Value |
| Intercept | -4.4712 | 1.0104 | 19.5839 | <0.0001 |
| Service at ICU admission: | -1.7275 | 0.4272 | 16.3527 | <0.0001 |
| PO2 from initial blood gases | 0.5553 | 0.5298 | 1.0988 | 0.2945 |
| AGE | 0.0602 | 0.0151 | 15.8519 | <0.0001 |
| Cancer part of the present problem | 1.3949 | 0.6532 | 4.5607 | 0.0327 |

1. Present a table containing odds ratios, 95% confidence intervals and p-values for your final model from homework assignment 6.

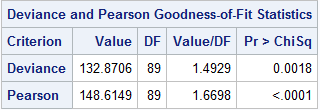
Table 1.2: Odds Ratios, 95% confidence intervals, and p-values for final Model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Comparison/Unit | OR | 95%CI | | P-Value |
| Service at ICU admission: | Surgical vs Medical | 0.178 | 0.077 | 0.411 | <0.0001 |
| PO2 from initial blood gases | <=60 vs >60 | 1.742 | 0.617 | 4.922 | 0.2945 |
| AGE | 10 | 1.825 | 1.357 | 2.454 | <0.0001 |
| Cancer part of the present problem | Yes vs No | 4.035 | 1.122 | 14.514 | 0.0327 |

1. Determine the overall goodness-of-fit of your model.

* 1. Use the deviance and Pearson GOF tests

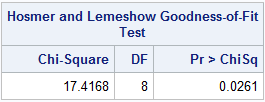
Table 2.1: Table Outlining the Deviance and Pearson Goodness-of-Fit Statistics



Based on the p-values from the Deviance and Pearson GOF test, this model is NOT a good fit. Better fitting model have a higher p-value. >0.25 would serve as good evidence of a better fit model, which we do not experience in this statistical test.

* 1. Use the Hosmer-Lemeshow test

Table 2.2: Table Outlining the Hosmer and Lemeshow Goodness of Fit Test



Based on the p-values from the Hosmer-Lemeshpw GOF test, this model is NOT a good fit. Better fitting model have a higher p-value. >0.25 would serve as good evidence of a better fit model, which we do not experience in this statistical test.

* 1. Use the Osius-Rojek test

Table 2.3: p-value for the Osius Rojek Test



Based on the p-values from the Osius-Rojek GOF test, this model is NOT a good fit. Better fitting model have a higher p-value. >0.25 would serve as good evidence of a better fit model, which we do not experience in this statistical test.

* 1. Describe which tests are and which tests are not appropriate and explain why.

**Appropriate:** Deviance and Pearson Goodness of fit tests.

**Inappropriate:** Hosmer-Lemshow, Osius-Rojek

The Deviance and Pearson goodness of Fit tests are good for a couple of reasons:

1. J < N.
   1. Whenever the amount of covariate patterns we have (in this case, 94) is less than the number of individuals in the study (ICU dataset n=200), then the pearson and deviance GOF tests are the best to determine if the model is able to be fit.
2. The Deviance an Pearson Goodness of fit Test are reasonably better with a smaller sample size (200 being considered a relatively small sample size).
3. Determine if your model is fitting the tails correctly.

Table 3.1: Significance tests for the upper (z1\_j) and lower tail (z2\_j) from the Stukel Test.



There is evidence that the shape of the lower tail may be modeled inadequately by this logistic model (p=0.0781). Although, the upper tail appears to be fitting correctly (p=0.3837).

Table 3.2: Overall p-value for tail fitting in the multivariate logistic regression model from the Stukel Test.



The overall p-value suggests evidence of model fit. But looking at the lower and upper tails, it is unclear. More than likely it is not because of the previous goodness of fit tests conducted earlier in this homework which suggest evidence of lack of model fit.

SAS CODE

libname sdat 'C:\Users\ndyet\_000\Desktop\Class Folders\Spring 2016\ERHS 642\Data';

**data** ICU\_altered; set sdat.ICU\_altered;

if race=**1** then do; r1=**0**; r2=**0**; end;

else if race=**2** then do; r1=**1**; r2=**0**; end;

else if race=**3** then do; r1=**0**; r2=**1**; end;

if **16**<= SYS <**110** then SYSa=**0**;

else if **110**<= SYS <**150** then SYSa=**1**;

else if SYS >= **150** then SYSa=**2**;

**run**;

\*\* Final model from chapter 4\*\*;

**proc** **logistic** descending data=ICU\_altered;

model STA=SER PO2 age CAN;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\* Pearson chi-square, deviance and Hosmer-Lemeshow test \*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **logistic** descending data=ICU\_altered;

model STA=SER PO2 age CAN

/scale=n aggregate lackfit;

**run**;

\*"/scale=n aggregate lackfit" gives you chi square, H-L and deviance, and remember H-L High p-value=good;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\* Osius-Rojek test \*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* If interaction terms are in the model, create the interaction terms ;

\* If categorical variables with more than 2 categories are in the model;

\* create design variables ;

\* Sort new data set by model covariates;

**proc** **sort** data=ICU\_altered; \*\*\*\* for a different data set change independent variable names in by statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

by SER PO2 age CAN;

**run**;

\* For each covariate pattern, j, save m\_j= # with covariate pattern j and;

\* y\_j = # with outcome=1 in covariate pattern j ;

**proc** **means** n sum noprint data=ICU\_altered; \*\*\*\* for a different data set change independent variable names in by statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\* and outcome variable name in var statement \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

by SER PO2 age CAN;

var STA; output out=jdat n=m\_j sum=y\_j; \*<--- indicate outcome;

**run**;

\*LOOK ABOVE! after "var" is where your outcome should be;

\* Run proc logistic for covariate patterns rather than individuals ;

\* outcome=y\_j / m\_j (not 0 or 1) , save fitted values ;

**proc** **logistic** noprint descending data=jdat; \*\*\*\* for a different data set change independent variable names in model statement. keep jdat as jdat. do NOT change this\*\*\*\*\*;

model y\_j/m\_j= SER PO2 age CAN;

output out=pdat p=p\_j;

**run**;

\* Create v\_j, c\_j, the chi-square terms and the terms in the sum in A;

**data** pdat; set pdat;

v\_j=m\_j\*p\_j\*(**1**-p\_j); c\_j=(**1**-**2**\*p\_j)/v\_j; chisq\_j=(y\_j-m\_j\*p\_j)\*\***2**/v\_j;

m\_j\_inv=**1**/m\_j;

**run**;

\* Create and save chi-square & sum for A ;

\* Perform weighted linear regression, save SS ;

\* Calculate RSS,A,z & p-val for z ;

**proc** **means** sum noprint data=pdat;

var chisq\_j m\_j\_inv; output out=cdat sum=chisq m\_inv; **run**;

**proc** **reg** noprint data=pdat outest=ss; \*\*\*\* for a different data set change independent variable names in model statement \*\*\*\*\*;

model c\_j=SER PO2 age CAN;

weight v\_j; **run**;

**data** zdat; merge cdat (keep=\_freq\_ chisq m\_inv) ss (keep=\_rmse\_);

rss=(\_freq\_-**4**-**1**)\*\_rmse\_\*\***2**; A=**2**\*(\_freq\_-m\_inv); \*\*\*\* for a different data set change 8 to number of variables in the model \*\*\*\*\*;

z=(chisq-(\_freq\_-**4**-**1**))/sqrt(A+rss); z=abs(z); \*\*\*\* for a different data set change 8 to number of variables in the model \*\*\*\*\*;

pval=(**1**-probnorm(z))\***2**;

**run**;

**proc** **print** noobs data=zdat; var pval; **run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\* Stukel test of logistic regression model assumption \*\*\*;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

ODS trace on;

ODS output GlobalTests=gt1;

**proc** **logistic** descending data=ICU\_altered; \*\*\*\* for a different data set change outcome and independent variable names \*\*\*\*\*;

\*\*\*\* in model statement \*\*\*\*\*;

model STA=SER PO2 age CAN;

output out=pdat2 xbeta=g\_j p=p\_j;

**run**;

**data** pdat2;

set pdat2;

if p\_j>=**0.5** then ind1=**1**; else ind1=**0**;

if p\_j< **0.5** then ind2=**1**; else ind2=**0**;

z1\_j=**0.5**\*g\_j\*\***2**\*ind1;

z2\_j=-**0.5**\*g\_j\*\***2**\*ind2;

**run**;

ODS output GlobalTests=gt2;

**proc** **logistic** descending data=pdat2 ; \*\*\*\* for a different data set change outcome and independent variable names \*\*\*\*\*;

\*\*\*\* in model statement but keep z1\_j and z2\_j \*\*\*\*\*;

model STA=SER PO2 age CAN z1\_j z2\_j;

**run**;

**data** pval;

merge gt1(rename=(ChiSq=ChiSq1))

gt2(rename=(ChiSq=ChiSq2));

if \_N\_=**1**;

drop Test df ProbChisq;

lr=ChiSq2-ChiSq1;

pval=(**1**-probchi(lr,**2**));

**run**;

**proc** **print** noobs data=pval; var pval; **run**;