Alla Topp

Logistic regression

Case: An investigator wants to study the relationship between age and presence or absence of coronary heart disease.

Patient	Age	Coronary Heart Disease
1	25	0
2	26	0
3	28	1
4	30	0
5	31	0
6	32	0
7	34	1
8	35	0
9	36	1
10	37	0
11	39	0
12	40	1
13	50	1
14	51	1
15	52	1
16	53	0
17	54	1
18	55	1
19	56	0
20	57	1
21	58	1
22	59	1
23	60	1

State your hypothesis and other assumptions.

H0: There is no association between coronary heart disease presence or absence and age (the odds ratio is equal to 1).

H1: There is association between coronary heart disease presence or absence and age (the odds ratio is equal to 1).

Show the command(s) and the corresponding results, Interpret the (age) coefficient and conclude the result

Alla Topp

```
> fit <- glm(CHD ~ Age, data = SomeDataSet, family = binomial(logit))
> summary(fit)# display results

Call:
glm(formula = CHD ~ Age, family = binomial(logit), data = SomeDataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9136  -0.8362   0.5120   0.7284   1.7253

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1212     1.9384  -2.126   0.0335 *
Age           0.1032     0.0451   2.288   0.0222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31.492  on 22  degrees of freedom
Residual deviance: 24.844  on 21  degrees of freedom
AIC: 28.844

Number of Fisher Scoring iterations: 4
```

The coefficients table gives the estimate values for the coefficients, or the betas, for our logistic regression model. We also see that our variable has one star, meaning that it’s significant in our model. We can reject the null hypothesis and state that the coefficient of age is significant because our p value is less than 0.05. We can say that there is a relationship between CHD and age. I assume that there is a presence in CHD based on age based on positive value of estimate std (0.1032). AIC value is a measure of the quality of the model and is like Adjusted R-squared in that it accounts for the number of variables used compared to the number of observations. A low AIC is desirable, which we have. Also, we can see a quite difference between null deviance and residual deviance, meaning the model is a good fit.

Compare the null model with your fit model. Does the age improve the model?

> anova(...., test='Chisq')

```
> #4 Compare the null model with your fit model
> anova(fit, test='Chisq') #compare nested models
Analysis of Deviance Table

Model: binomial, link: logit

Response: CHD

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    22     31.492
Age   1   6.6482        21     24.844 0.009926 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Alla Topp

Here is the p-value < 0.05, there is a high evidence to reject the null hypothesis. This means that the fitted model (with age) reduces the deviance and this reduction is significant. We can say that the age improves the model.

Use the model to predict the probability of a random age (one value or a range of values) for the potential of coronary heart disease.

> predict(....., type = 'response')

With predict function we are trying to predict if the age affects the occurrence of CHD. We can see that each value is getting higher. I assume that it tells us that chances of getting CHD when you get older is higher based on the analysis.

```
> predict(fit, type="response") # predicted values
        1         2         3         4         5         6         7         8         9        10        11        12 
0.1762506 0.1917325 0.2257518 0.2638369 0.2843566 0.3058093 0.3512728 0.3751276 0.3996044 0.4245932 0.4756169 0.5013895 
       13        14        15        16        17        18        19        20        21        22        23 
0.7383164 0.7577538 0.7761852 0.7935961 0.8099842 0.8253575 0.8397331 0.8531359 0.8655973 0.8771536 0.8878449 
```