DEPARTMENT OF STATISTICS

Stats 762: Topics in Regression Modelling

Model answers for Term Test: Thursday, October 7 2010

1. *Under what circumstances can residuals in a logistic regression analysis be interpreted in the same way as residuals in a regression having a continuous response?*

   When the number of covariate patterns is small and there are a reasonable number of observations for each covariate pattern.

2. *Suppose* **`q2.glm`** *is the result of fitting a logistic regression using the glm function. Explain the difference between the results of the two bits of code*

   (a) `predict(q2.glm)` *and*
   (b) `predict (q2.glm, type ="response")`

   (a) Calculates the fitted log-odds while (b) calculates the fitted probabilities.

3. *The data for this question come from a study that investigated the effect of insulin on laboratory mice. The response was whether or not the mice had convulsions when given insulin. We are interested in modelling how the proportion of mice with convulsions varies with the dose applied.*

Dose (mg)	*Number with convulsions*	*Number of mice*
3.4	***0***	***33***
5.2	***5***	***32***
7.0	***11***	***38***
8.5	***14***	***37***
10.5	***18***	***40***
13.0	***21***	***37***
18.0	***23***	***31***
21.0	***30***	***37***
28.0	***27***	***30***

A logistic model `cbind(r, n-r) ~dose` *was fitted, resulting in the output below*

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.44388    0.30018  -8.141 3.91e-16 ***
dose         0.19295    0.02351   8.208 2.25e-16 ***
```

According to the fitted model, what is the effect on the odds of convulsions of increasing the dose from 7mg to 10 mg?

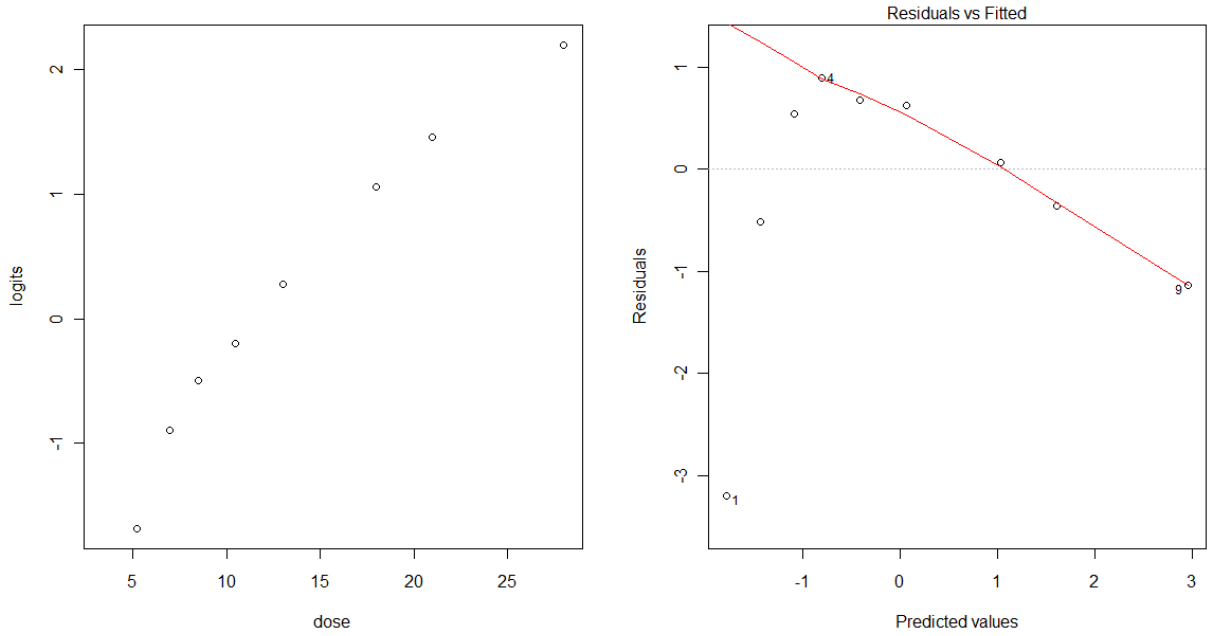
odds of convulsions at 10 mg / odds of convulsions at 10 mg
= exp(-2.44388 + 10*0.19295)/ exp(-2.44388 + 7*0.19295)
= exp(3*0.19295)
= 1.783986

Thus, the odds are increased by about 78%

4. *Use the fitted model to estimate the dose at which 50% of the mice have convulsions*.

A probability of 0.5 is equivalent to a log-odds of 0. Thus we need to solve the equation
0 = -2.44388 + x*0.19295. This gives x = 2.44388/0.19295 = 12.66587
Thus, the dose required is about 12.66 mg.

5. *For the regression in Q3, a plot of log(r/(n-r) versus dose, and a plot of deviance residuals versus predicted values is shown below. Do these plots indicate any problems with the regression? If so, what is wrong and what would you do to fix the problem?*



The logits are not a linear function of dose, so dose should be transformed.

6. *In the logistic regression fitted in Q3, the maximum value of the unrestricted (i.e. maximal) log-likelihood is -159.5074, the maximum value of the logistic log-likelihood is –166.4280. What is the value of the residual deviance?*

The deviance is D = 2(-159.5074 – (–166.4280)) = 13.8412.

7. *Suppose we fit a logistic model. Briefly explain how we can assess the goodness of fit of the model in the case of (a) grouped data, where there are a reasonably large number of observations for each covariate pattern, and (b) for ungrouped data, where there is only one observation for each covariate pattern.*

For grouped data, use the residual deviance and residual plots. For ungrouped, use the Hosmer-Lemeshow statistic

8. *The table overleaf refers to the 1980 US presidential election. The entries r in the table give the number of respondents in a survey who voted for Ronald Reagan in the election, out of n total respondents, for each rating-race combination. The rating is a rating of political conservatism, with 1 representing extremely liberal and 7 extremely conservative. Thus, there were a total of 13 respondents classified as White and "extremely liberal", of which 1 voted for Regan.*

```
          Rating
          1  2   3   4   5   6  7
White r   1 13  44 155  92 100 18
      n  13 70 115 301 153 141 26
Black  r  0  0   2   1   0   2  0
      n   6 16  25  32   8   9  4
```

```
r = c(1, 13, 44, 155, 92, 100, 18, 0, 0, 2, 1, 0, 2, 0)
n = c(13, 70, 115, 301, 153, 141, 26, 6, 16, 25, 32, 8, 9, 4)
rating = factor(rep(1:7,2))
race = factor(rep(c("Black","White"), each=7))
pol.glm = glm(cbind(r, n-r)~rating*race, family=binomial)
anova(pol.glm)
```

```
Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            13    185.157
rating       6  102.861         7     82.296    <2e-16 ***
race         1   77.335         6      4.961    <2e-16 ***
rating:race  6    4.961         0      0.000    0.5488
```

Explain why the model `cbind(r, n-r)~rating + race` *is indicated by this output.*

Because the p-value for the interaction term (0.5488) indicates that the interaction hypothesis is not rejected.

9. *The model* `cbind(r, n-r)~rating + race` *was fitted , producing the output below. Based on this output, what is the estimated probability that a Black person having rating 4 will vote for Reagan?*

```
> pol2.glm = glm(cbind(r, n-r)~rating+race, family=binomial)
> summary(pol2.glm)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.4012     1.1347  -4.760 1.94e-06 ***
rating2       1.0176     1.0828   0.940  0.34734
rating3       2.0774     1.0555   1.968  0.04905 *
rating4       2.5640     1.0449   2.454  0.01413 *
rating5       2.9087     1.0514   2.766  0.00567 **
rating6       3.4370     1.0547   3.259  0.00112 **
rating7       3.2511     1.1153   2.915  0.00356 **
raceWhite     2.8867     0.4707   6.133 8.62e-10 ***
```

The log odds of voting for Reagan are -5.4012+2.5640= -2.8372, so the probability is exp(-2.8372)/(1+exp(-2.8372))= 0.0553.

10. How could you check to see if these data were over-dispersed?

Refit the model using family=quasibinomial and examine the value of the dispersion parameter.