MATH 424: Homework Chapter 9: Logistic Regression

Due on Wednesday, December 6, 2017

Kafai 11:10am

Jonathan Dombrowski

Contents

Q 24

Flight response of geese. Offshore oil drilling near an Alaskan estuary has led to increased air traf- ficmostly large helicoptersin the area. The U.S. Fish and Wildlife Service commissioned a study to investigate the impact these helicopters have on the flocks of Pacific brant geese, which inhabit the estuary in Fall before migrating (Sta- tistical Case Studies: A Collaboration between Academe and Industry, 1998). Two large heli- copters were flown repeatedly over the estuary at different altitudes and lateral distances from the flock. The flight responses of the geese (recorded as low or high), altitude (x 1 = hundreds of meters), and lateral distance (x 2 = hundreds of meters) for each of 464 helicopter overflights were recorded and are saved in the PACGEESE file. (The data for the first 10 overflights are shown in the table, p. 503.) MINITAB was used to fit the logistic regression model = 0 + 1 x 1 + 2 x 2 , where y = 1 if high response, 0 if low response, = P(y = 1), and = ln[/(1)]. The result- ing printout is shown above.
(a) Is the overall logit model statistically useful for predicting geese flight response? Test using = .01.
(b) Conduct a test to determine if flight response of the geese depends on altitude of the heli- copter. Test using = .01.
(c) Conduct a test to determine if flight response of the geese depends on lateral distance of helicopter from the flock. Test using = .01.
(d) Predict the probability of a high flight response from the geese for a helicopter flying over the estuary at an altitude of x 1 = 6 hundred meters and at a lateral distance of x 2 = 3 hundred meters.

a

$$H_0 = \beta_1 = \beta_2 = 0$$

$$H_a = \beta_1, \beta_2 \neq 0$$

We can compare the Chi squared values and their respective d.f.'s by invoking
'1-pchisq(model$null.deviance - model$deviance, model$df.null - model$df.residual)'
In the same motion as the F-test as we would do for a linear regression: taking the ratio of two Chi squared values and exracting a p-value from that.
We can extract a p-value = 0. p ¡ α = 0.01, therefore we can reject H_0 and state that the model is statistically useful for predicting geese flight response.

b

$$H_0 = \beta_1 = 0$$

$$H_a = \beta_1 \neq 0$$

by looking at the model summary, we can see that the z-score for this test is z=2.914 , with a corresponding p-value of 0.00357. We can conclude that p ¡ α and therefore that the flight response does in fact depend on the altitude of the helicopter.

c

$$H_0 = \beta_2 = 0$$

$$H_a = \beta_2 \neq 0$$

by looking at the model summary, we can see that the z-score for this test is z=-10.625 , with a corresponding p-value of less than 2×10^{-16}. We can conclude that p ¡ α and therefore that the flight response does in fact depend on the lateral distance of the helicopter.

d

For

$$x_1 = 6, x_2 = 3$$

where both are measured in hundreds of meters, the predicted probability from the logistic model

$$\pi^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

using the R commands:
¿ newdata = data.frame(ALTITUDE=6, LATERAL=3)
¿ predict(model,newdata,type="response")
is 0.946. From this we can conclude that there is a 94.6% chance that the geese have a high flight response when the Altitude is held at a fixed 600m, and the Lateral distance is held at a fixed 300m based on the data reported.

Q 26

Groundwater contamination in wells. Many New Hampshire counties mandate the use of reformulated gasoline, leading to an increase in groundwater contamination. Refer to the Environmental Science and Technology (January 2005) study of the factors related to methyl tert-butyl ether (MTBE) contamination in public and private New Hampshire wells, Exercise 6.11 (p. 343). Data were collected for a sample of 223 wells and are saved in the MTBE file. Recall that the list of potential predictors of MTBE level include well class (public or private), aquifer (bedrock or unconsolidated), pH level (standard units), well depth (meters), amount of dissolved oxygen (mil ligrams per liter), distance from well to nearest fuel source (meters), and percentage of adjacent land allocated to industry. For this exercise, consider the dependent variable y = 1 if a detectible level of MTBE is found, 0 if the level of MTBE found is below limit. Using the independent variables indentified in Exercise 6.11, fit a logistic regres sion model for the probability of a detectible level of MTBE. Interpret the results of the logistic regression. Do you recommend using the model? Explain.

a

The model in question is defined as follows:

$$\pi^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

```
        Call:
glm(formula = detet ~ wellclass + aquifer + ph + welldepth +
    do2 + d2f + pind, family = binomial(link = "logit"))
```

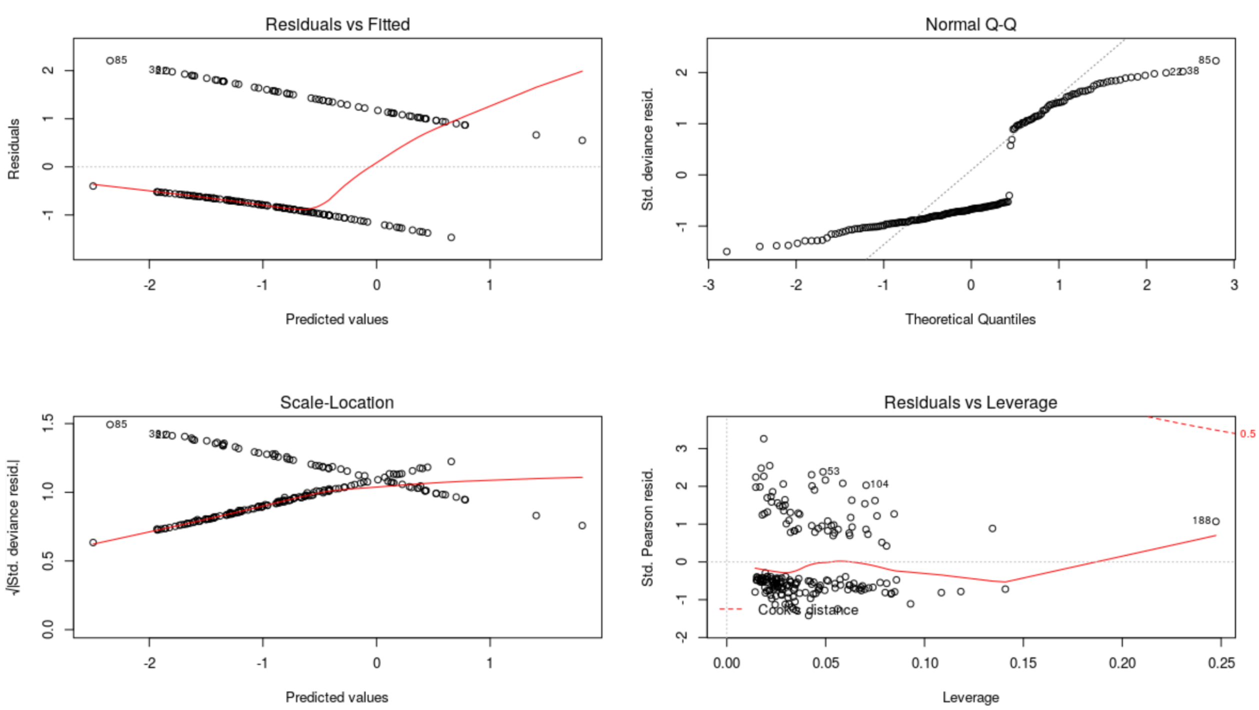
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4664  -0.8663  -0.6549   1.0395   2.2075

Coefficients:
                   Estimate Std. Error z value Pr(>z)
(Intercept)        1.1713423  1.7544128   0.668   0.5044
wellclassPublic    0.8066518  0.3807905   2.118   0.0341 *
aquiferUnconsoli  -0.2693525  0.6760558  -0.398   0.6903
ph                -0.4098715  0.2328449  -1.760   0.0784 .
welldepth          0.0084778  0.0034056   2.489   0.0128 *
do2                0.0062126  0.0742396   0.084   0.9333
d2f               -0.0001209  0.0001642  -0.736   0.4617
pind               0.0234640  0.0317899   0.738   0.4605
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 242.21  on 190  degrees of freedom
Residual deviance: 221.13  on 183  degrees of freedom
  (32 observations deleted due to missingness)
AIC: 237.13

Number of Fisher Scoring iterations: 4
```



Looking at the graphs for this model from the R output, we can see that the residuals for both groups are

distributed normally, with none of them breaching 3 standards deviations, or the Cooks' distance. So the weight of any single point does not overly pull the model in any direction too far.
Continuing the analysis, we test the model for overall model adequacy.

$$H_0 = \beta_1 = \beta_2 = \ldots = \beta_7 = 0$$

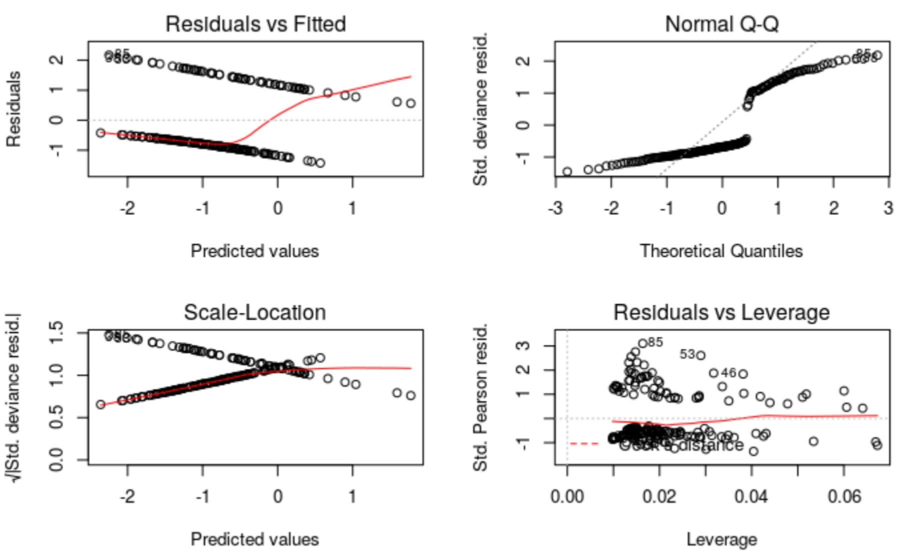
$$H_a = \text{at least one } \beta_i (i = 0, 7) \neq 0$$

To do this we use the same method as in Question 24. From our χ^2 testing, we get a p-value = 0.00365, which: $p - value < \alpha = 0.5$, so we can reject the null hypothesis and state that the model is statistically useful for predicting the probability of detection in MTBE levels.
After globally validating the model, we can then move to see if each of the β's are statistically significant. For i = 1 to 7 :

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

We can refer to the R printout to see the p-values associated with the z-scores for the individual betas. After doing a backwards variable selection based on the p-values, we arrive at a model predicting the Detection status of MTBE with the WellClass, pH level, and WellDepth. The p-values for the respective predictors are 0.00575, 0.05476, 0.00187. The pH level was the only one to come close to the default $\alpha = 0.05$. Testing overall model adequacy again, we use the χ^2 testing method form Q 24, we arrive at a p-value of 0.000258. Comparing this to the previous model's p-value of 0.00365. Following the manual variable selection, we have improved the p-value by a factor of 10.



Looking at the residuals quickly, after model reformation, we see that nothing has changed in respect to the status of normality or leverage.
Depending on the tolerance of the client requesting the model, I would also reccommend removing the pH

term as the p-value was just above 0.05. Depending on the desire of the company, I would reccommend either this model, or this model with the pH value removed.