**Section 5.2 – Tools to assess model fit (continued, part 3)**

Goodness of fit (GOF)

Residual measures examine how well a model fits an individual observation. GOF statistics evaluate the model for all observations at once.

Two commonly used GOF statistics are the Pearson and residual deviance statistics:

Pearson statistic: 

Residual deviance statistic: D = -2log(Λ) for comparing the model of interest to the saturated model

Questions:

* Why do you think these statistics would be useful to examine?
* The statistics are used for a GOF hypothesis test. What are the hypotheses in words, rather than symbols?
* What types of values for X2 and D would lead to a rejection of the null hypothesis?

These statistics have approximate  distributions for “large” samples, where  denotes the number of regression parameters (p+1: β0, …, βp) estimated in the model of interest. However, this result is only valid under the assumption that the number of unique sets of explanatory variable combinations in the data is fixed as the sample size increases.

As an informal alternative, we recommend calculating “Deviance/df”, i.e., . If the model is reasonable, this numerical value should not be too far from 1. How far away from 1 is still o.k.? General guidelines are

* Potential problem: 
* Poor fit: 

Where do these guidelines come from?

Note that . *If* D was well approximated by a , E(D) ≈  and Var(D) ≈ . Using ±2 standard deviations from the mean, we have 1 ± 2. Note that first and second moments should be better approximated than the entire distribution.

There are other measures of goodness-of-fit too. These are described in my book.

Influence

An observation is influential to a model’s fit if removing or changing the observation results in a significant change in the regression parameter estimates or estimated response values (). There are a number of influence measures that can be calculated for an observation. They all focus on removing the observation from the data set and then examining how a particular statistic changes. Because removing each observation one at a time and re-estimating a model can be time consuming, approximations have been developed for these measures to avoid needing to re-estimate models.

Cook’s distance – This is an overall measure of how all parameter estimates would change if the mth observation is removed from the data set. Cook’s distance is



and it is always greater than 0. Those observations with CDm values much larger than the rest should be investigated further for potentially being influential. Others use the general guideline of CDm > 1 may have high influence and CDm > 4/M may have moderate influence.

Change in Pearson GOF statistic – This is a measure of how much X2 will change if the mth observation is removed from the data set. The “delta X-square” statistic is



The statistic is just the square of the standardized Pearson statistic! A  approximation can be used with the statistic, but with the same caveats as discussed earlier for rm and a standard normal distribution approximation. For this reason, I recommend simply looking at  > 4 and  > 9 as thresholds.

Change in the residual deviance statistic – This is a measure of how much the residual deviance will change if the mth observation is removed from the data set. The “delta D” statistic is



The same approximation as used with  can be used here.

Hat matrix diagonal value (leverage) – This measure alone is sometimes used with normal linear regression models to identify outlying observations (in terms of their explanatory variable values), which may subsequently be influential. For GLMs, interpreting the hm becomes more difficult because a large value may not mean being outlying with respect to the explanatory variables (see discussion in book). Overall, some individuals use the guidelines that values larger than 2p/M indicate “moderately high leverage,” and values larger than 3p/M indicate high leverage. Note that having a high amount of leverage is not necessarily bad. This is why I focus on this measure MUCH less than the others.

Calculation in R – Use the model fit object obtained from glm() with the following functions:

* hatvalues() calculates hm
* cooks.distance() calculates CDm
* rstandard() calculates  which then can be squared for  (type = "pearson")
* rstandard() calculates  (type = "deviance") which subsequently leads to the calculation of 

My book discusses a general function written for the book named glmInflDiag() (glmDiagnostics.R) which automatically calculates these measures and produces plots. This function works well for GLMs in general. The last section of Chapter 5 discusses an alternative function specifically for logistic regression models named examine.logistic.reg(). We will use this function shortly.

Example: Plot with about 50 EVPs included

One plot produced by examine.logistic.reg() includes the  values on the y-axis and the  values on the x-axis.



The above is a commonly seen pattern for these plots, especially when many nm values are small. The pattern itself is not of concern.

As defined earlier,



There are some large  values in the plot:

* When  is close to 0, these EVPs correspond to those with wm > nm.
* When  is close to 1, these EVPs correspond to those with a wm < nm.

If the  approximation is poor (when nm is relatively small), these cases will occur and may not be something to be concerned about. The individual EVPs should still be examined to determine if there is a reason beyond simply the discrete aspects of the response.

Example: Placekicking (Placekick\_NotInBook.R, Examine.logistic.reg.R, Placekick.csv)

> placekick <- read.csv(file = "C:\\data\\placekick.csv")

> w <- aggregate(x = good ~ distance, data =

 placekick, FUN = sum)

> n <- aggregate(x = good ~ distance, data =

 placekick, FUN = length)

> w.n <- data.frame(distance = w$distance, success =

 w$good, trials = n$good, prop = round(w$good/n$good,4))

> mod.fit.bin <- glm(formula = success/trials ~ distance,

 weights = trials, family = binomial(link = logit), data

 = w.n)

> # summary(mod.fit.bin)

> source(file = "C:\\Rprograms\\Examine.logistic.reg.R")

> save.info1 <- examine.logistic.reg(mod.fit.obj = mod.fit.bin, identify.points = TRUE)



Plots:

* (1,1): This is the same plot as we saw before
* (1,2): Observation #3 is EXTREME on the plot! Why do you think this occurs?
* (2,1): The y-axis numerical values are the squares of the y-axis numerical values on the (1,1) plot. What makes this plot helpful then is that the size of the plotting symbol is proportional to nm.
* (2,2): The same plot as in (2,1), but now the plotting symbol is proportional in size to Cook’s distance.
* Deviance/df is given below the plot

What do you think of the model?

When there are a few observations with nm or CDm values that are much larger than the rest, this can make examining the size of the plotting symbol not helpful for the other observations. For this reason, I have included a way to rescale the numerical values used as the plotting symbol size. Below are some examples.

> save.info2 <- examine.logistic.reg(mod.fit.obj =

 mod.fit.bin, identify.points = TRUE, scale.n = sqrt,

 scale.cookd = sqrt)



> one.fourth.root <- function(x) {

 x^0.25

 }

> one.fourth.root(16) # Example

[1] 2

> save.info3 <- examine.logistic.reg(mod.fit.obj = mod.fit.bin, identify.points = TRUE, scale.n = one.fourth.root, scale.cookd = one.fourth.root)



For those observations identified in the plots, it is often helpful to list out all their explanatory variable and response values with their residual and influence measures. Because there are not too many observations, I did this for all observations:

> data.frame(w.n[,-4], prop = round(w.n[,4],2),

 pi.hat = round(mod.fit.bin$fitted.values,2),

 std.Pe = round(save.info1$stand.resid,2),

 tail.p = round(save.info1$tail.prob,2),

 dXsq = round(save.info1$deltaXsq,2),

 cook = round(save.info1$cookd,2))

 distance success trials prop pi.hat std.Pe tail.p dXsq cook

1 18 2 3 0.67 0.98 -3.58 0.07 12.78 0.02

2 19 7 7 1.00 0.97 0.43 0.83 0.19 0.00

3 20 776 789 0.98 0.97 3.63 0.02 13.16 13.16

4 21 19 20 0.95 0.97 -0.45 0.48 0.20 0.00

5 22 12 14 0.86 0.96 -2.15 0.09 4.62 0.03

6 23 26 27 0.96 0.96 0.09 0.67 0.01 0.00

7 24 7 7 1.00 0.95 0.58 0.72 0.33 0.00

8 25 12 13 0.92 0.95 -0.44 0.49 0.19 0.00

9 26 8 9 0.89 0.94 -0.72 0.41 0.52 0.00

10 27 24 29 0.83 0.94 -2.48 0.03 6.13 0.09

11 28 20 22 0.91 0.93 -0.39 0.46 0.16 0.00

12 29 16 17 0.94 0.92 0.29 0.62 0.08 0.00

13 30 12 14 0.86 0.91 -0.76 0.34 0.58 0.00

14 31 10 11 0.91 0.90 0.05 0.67 0.00 0.00

15 32 23 30 0.77 0.89 -2.30 0.03 5.28 0.08

16 33 20 21 0.95 0.88 1.01 0.27 1.01 0.01

17 34 16 19 0.84 0.87 -0.37 0.46 0.13 0.00

18 35 12 14 0.86 0.86 0.01 0.62 0.00 0.00

19 36 18 22 0.82 0.84 -0.31 0.47 0.09 0.00

20 37 22 29 0.76 0.83 -0.97 0.23 0.94 0.02

21 38 23 28 0.82 0.81 0.18 0.55 0.03 0.00

22 39 22 28 0.79 0.79 -0.06 0.55 0.00 0.00

23 40 13 19 0.68 0.77 -0.91 0.26 0.83 0.01

24 41 7 10 0.70 0.75 -0.36 0.48 0.13 0.00

25 42 20 27 0.74 0.73 0.16 0.54 0.03 0.00

26 43 12 22 0.55 0.70 -1.67 0.08 2.79 0.07

27 44 13 18 0.72 0.68 0.40 0.46 0.16 0.00

28 45 11 18 0.61 0.65 -0.39 0.44 0.15 0.00

29 46 11 14 0.79 0.63 1.25 0.17 1.57 0.04

30 47 14 23 0.61 0.60 0.09 0.56 0.01 0.00

31 48 10 18 0.56 0.57 -0.15 0.53 0.02 0.00

32 49 8 12 0.67 0.54 0.88 0.29 0.77 0.02

33 50 10 19 0.53 0.52 0.10 0.55 0.01 0.00

34 51 11 15 0.73 0.49 2.00 0.05 3.98 0.17

35 52 5 13 0.38 0.46 -0.55 0.40 0.30 0.01

36 53 5 9 0.56 0.43 0.79 0.33 0.62 0.02

37 54 1 7 0.14 0.40 -1.43 0.16 2.04 0.05

38 55 2 3 0.67 0.37 1.06 0.32 1.12 0.01

39 56 1 1 1.00 0.35 1.37 0.35 1.89 0.01

40 59 1 1 1.00 0.27 1.63 0.27 2.67 0.01

41 62 0 1 0.00 0.21 -0.52 0.79 0.27 0.00

42 63 0 1 0.00 0.19 -0.49 0.81 0.24 0.00

43 66 0 1 0.00 0.14 -0.41 0.86 0.17 0.00

When there is only one explanatory variable, another good plot to examine is what we saw in Chapter 2:



Remember that the very large distance placekicks do not have many observations.