**Section 5.2 – Tools to assess model fit**

This was a difficult section for us to write for the following reasons:

* The generalized linear models (GLMs) of Chapters 2 and 4 largely fit within the same framework, so we wanted to discuss them at the same time. This led us to use a little different notation to encompass these GLMs.
* There are enough differences among the GLMs of Chapters 2 and 4 that some differentiation is still needed at times.
* The multicategory models of Chapter 3 are not exactly GLMs. The goals to evaluate these models though are largely the same as for GLMs. Unfortunately, evaluation tools are much less developed and not implemented in statistical software packages. This led us to briefly talk about these models at the end of the section, separate from GLMs.

Therefore, the discussion next is for the GLMs presented in Chapters 2 and 4. Watch for the few notational changes.

Explanatory variable pattern (EVP) form

In the placekicking data set, each of the 1,425 rows represented one observation:

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

> head(placekick)

week distance change elap30 PAT type field wind good

1 1 21 1 24.7167 0 1 1 0 1

2 1 21 0 15.8500 0 1 1 0 1

3 1 20 0 0.4500 1 1 1 0 1

4 1 28 0 13.5500 0 1 1 0 1

5 1 20 0 21.8667 1 0 0 0 1

6 1 25 0 17.6833 0 0 0 0 1

The good variable represents a Bernoulli response – 0 for a failure and 1 for a success. When examining how well a model for this type of response fits the data, it is often better to examine the data in explanatory variable pattern (EVP) form. This format has one row in the data set for each of the unique sets of explanatory variables.

We saw EVP form already in Chapter 2 when we converted the placekick data set (involving ONLY distance and good) to a binomial response format:

> 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$good, proportion = round(w$good/n$good,4))

> head(w.n)

distance success trials proportion

1 18 2 3 0.6667

2 19 7 7 1.0000

3 20 776 789 0.9835

4 21 19 20 0.9500

5 22 12 14 0.8571

6 23 26 27 0.9630

> tail(w.n)

distance success trials proportion

38 55 2 3 0.6667

39 56 1 1 1.0000

40 59 1 1 1.0000

41 62 0 1 0.0000

42 63 0 1 0.0000

43 66 0 1 0.0000

For example, there are 20 observations with distance = 21. Of the 20, 19 are successes. Remember that when we estimated a model to this form of the data, we obtained the same regression parameter estimates as with the original Bernoulli form of the data.

Why is EVP form better than the original form of the data?

Many of our statistical measures that examine how well the models of Chapter 2 fit the data are based on continuous distributional approximations. These approximations work better with the EVP format as you will soon see.

Before evaluating the fit of models from Chapter 2, we will always need to convert the data to EVP form first.

Residuals

Let w denote the observed count. Thus, this can be the usual count from Chapter 4, but also the number of successes for a binomial response.

A residual in its simplest form is

Observed value – Estimated value

The residuals for Chapters 2 and 4 are then

Chapter 2 models: The residual is w – n

Chapter 4 models: The residual is w – 

where these residuals are calculated for EVERY observation in a data set. For example, the residual for the placekicking data set at 21 yards is

19 – 20×0.9676 = –0.3519

where



using the model from Chapter 2.

Notational changes:

* To write these residuals without specifying the Chapter number, I will use  to denote the estimated value from the model. Thus,

Chapter 2 models: n will be denoted by 

Chapter 4 models:  will be denoted by 

The residual then simply becomes w – .

* There will be m = 1, …, M “observations” rather than i = 1, …, n observations in a data set. The reason for this notation is that it can get confusing with respect to the models of Chapter 2 due to the data perhaps being observed in a binary response format (like for the placekicking data). Thus, the residuals are



for m = 1, …, M. For the Chapter 2 models, there are nm Bernoulli responses or trials for each m = 1, …, M.

Pearson residuals

To determine if a residual is unusual in size (i.e., an outlier), we need to take into account the variability in the residual. This is partially done by calculating Pearson residuals:



where  is the estimated variance for Wm depending on the model:

Logistic – 

Poisson – 

For the distance of 21 yards, there were nm = 20 trials and wm = 19 successes observed. The Pearson residual is then



The calculation of Pearson residuals is simple, but they have one problem – the denominator fails to take into account the variation in the estimated value given in the numerator. Thus, a standardized Pearson residual is



where hm come from the “hat matrix”. If you had a full regression course, this matrix should be very familiar! If not (or if you need a refresher), below is an explanation of the hat matrix:

Let **X** be the matrix of the explanatory variable data values with 1’s in the first column. Create a diagonal matrix  with diagonal elements of  in the same order as the corresponding data listed in **X**. The hat matrix is **H** = **X**(**X**′**X**)-1**X**′. Note that this is essentially the same hat matrix used when fitting a regression model by weighted least squares.

Both em and rm have “approximate” standard normal distributions. Thus, values of em and rm outside of  could be considered to be outliers. HOWEVER, this approximation does not always work well. This approximation is discussed next for the Chapters 2 and 4 models:

**Chapter 2 models** – We are essentially making a normal approximation for binomial random variables. This approximation depends on nm being large relative to probability of success (say, πm). For example, consider the case with an estimated probability of success  = 0.9 for an EVP where wm = 1 success out of nm = 3 trials is observed. The Pearson residual is:



Using a strict normal approximation, one would think this observation is being poorly fit by the model. However, note that there are ONLY 4 possible values for em:

> w <- 0:3

> n <- 3

> pi.hat <- 0.9

> e <- round((w - n\*pi.hat)/sqrt(n\*pi.hat\*(1-pi.hat)), 2)

> prob.norm <- round(pnorm(q = e), 4)

> prob.bin <- round(pbinom(q = w, size = n, prob = pi.hat), 4)

> data.frame(w, e, prob.norm, prob.bin)

w e prob.norm prob.bin

After recording video: Suppose W is a binomial random variable. Then dbinomial() finds P(W = w) and pbinomial() finds P(W ≤ w). Because π is unknown, we are using  to approximate it.

1 0 -5.20 0.0000 0.001

2 1 -3.27 0.0005 0.028

3 2 -1.35 0.0885 0.271

4 3 0.58 0.7190 1.000

Thus, em is a DISCRETE random variable, and we used a CONTINUOUS random variable to make a judgment about it. The normal distribution approximation should not be expected to do well!

It is not just nm that one needs to examine. For example, if nm = 100 and  = 0.99, one can see that there are few likely possible values for wm in this case, so the normal approximation would likely not work well here too.

**Chapter 4 models** – We are essentially making a normal approximation for Poisson random variables. For this approximation to work,  cannot be extremely small (say, need  > 0.5). Note that when  is extremely small, then one will expect to observe wm values that are equal to 0 or 1.

Because of these normal distribution approximation problems, I recommend being careful with making judgments about observations. Using simple guidelines, like em or rm being outside of ±2 or ±3, can be used to provide an initial impression about whether an observation is an outlier. One should then follow-up with a closer look at quantities like nm to make a more informed judgment.

Comments:

* The standardized Pearson residuals are better than the Pearson residuals to examine. I discuss Pearson residuals here because they lead nicely into the standardized version. Still, there may be cases where Pearson residuals are much easier to calculate (outside of models in Chapters 2 and 4), so this is another reason why they are introduced here.
* In what situation would you expect nm ≈ 1 for m = 1, …, M? How does this affect the use of em and rm to judge the model’s fit?
* The residual deviance can be written as  where dm is the specific contribution by the mth observation. The deviance residual is then based on dm:



Thus, the residual deviance is

.

The standardized deviance residual is



* The residuals() function with type = "pearson" calculates Pearson residuals.
* The rstandard() function with type = "pearson" calculates the standardized Pearson residuals.