GLM Residuals and Diagnostics

Patrick Breheny

March 26

Introduction

- After a model has been fit, it is wise to check the model to see how well it fits the data
- In linear regression, these diagnostics were build around residuals and the residual sum of squares
- In logistic regression (and all generalized linear models), there are a few different kinds of residuals (and thus, different equivalents to the residual sum of squares)

"The" χ^2 test

- Before moving on, it is worth noting that both SAS and R report by default a χ^2 test associated with the entire model
- This is a likelihood ratio test of the model compared to the intercept-only (null) model, similar to the "overall F test" in linear regression
- This test is sometimes used to justify the model
- However, this is a mistake

“The” χ^2 test (cont’d)

- Just like all model-based inference, the likelihood ratio test is justified under the assumption that the model holds
- Thus, the F test takes the model as given and cannot possibly be a test of the validity of the model
- The only thing one can conclude from a significant overall χ^2 test is that, if the model is true, some of its coefficients are nonzero (is this helpful?)
- Addressing the validity and stability of a model is much more complicated and nuanced than a simple test, and it is here that we now turn our attention

Pearson residuals

- The first kind is called the *Pearson residual*, and is based on the idea of subtracting off the mean and dividing by the standard deviation
- For a logistic regression model,

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

- Note that if we replace $\hat{\pi}_i$ with π_i, then r_i has mean 0 and variance 1

Deviance residuals

- The other approach is based on the contribution of each point to the likelihood
- For logistic regression,

$$\ell = \sum_i \{y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)\}$$

- By analogy with linear regression, the terms should correspond to $-\frac{1}{2}r_i^2$; this suggests the following residual, called the *deviance residual*:

$$d_i = s_i\sqrt{-2\{y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)\}},$$

where $s_i = 1$ if $y_i = 1$ and $s_i = -1$ if $y_i = 0$

Deviance and Pearson's statistic

- Each of these types of residuals can be squared and added together to create an RSS-like statistic
- Combining the deviance residuals produces the *deviance*:

$$D = \sum d_i^2$$

which is, in other words, -2ℓ

- Combining the Pearson residuals produces the *Pearson statistic*:

$$X^2 = \sum r_i^2$$

Goodness of fit tests

- In principle, both statistics could be compared to the χ^2_{n-p} distribution as a rough goodness of fit test
- However, this test does not actually work very well
- Several modifications have been proposed, including an early test proposed by Hosmer and Lemeshow that remains popular and is available in SAS
- Other, better tests have been proposed as well (an extensive comparison was made by Hosmer *et al.* (1997))

The hat matrix for GLMs

- As you may recall, in linear regression it was important to divide by $\sqrt{1-H_{ii}}$ to account for the leverage that a point had over its own fit
- Similar steps can be taken for logistic regression; here, the projection matrix is

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2},$$

where $\mathbf{W}^{1/2}$ is the diagonal matrix with $\mathbf{W}_{ii}^{1/2} = \sqrt{w_i}$

Properties of the hat matrix

- In logistic regression, $\hat{\boldsymbol{\pi}} \neq \mathbf{H}\mathbf{y}$ – no matrix can satisfy this requirement, as logistic regression does not produce linear estimates
- However, it has many of the other properties that we associate with the linear regression projection matrix:
  - $\mathbf{H}\mathbf{r} = \mathbf{0}$
  - $\mathbf{H}$ is symmetric
  - $\mathbf{H}$ is idempotent
  - $\mathbf{H}\mathbf{W}^{1/2}\mathbf{X} = \mathbf{W}^{1/2}\mathbf{X}$ and $\mathbf{X}^T\mathbf{W}^{1/2}\mathbf{H} = \mathbf{X}^T\mathbf{W}^{1/2}$

  where $\mathbf{r}$ is the vector of Pearson residuals

Standardized residuals

- The diagonal elements of $\mathbf{H}$ are again referred to as the *leverages*, and used to standardize the residuals:

$$r_{si} = \frac{r_i}{\sqrt{1 - H_{ii}}}$$
$$d_{si} = \frac{d_i}{\sqrt{1 - H_{ii}}}$$

- Generally speaking, the standardized deviance residuals tend to be preferable because they are more symmetric than the standardized Pearson residuals, but both are commonly used

Leave-one-out diagnostics

- You may recall that in linear regression there were a number of diagnostic measures based on the idea of leaving observation i out, refitting the model, and seeing how various things changed (residuals, coefficient estimates, fitted values)
- You may also recall that for linear regression, it was not actually necessary to refit the model n times; explicit shortcuts based on $\mathbf{H}$ were available
- The same idea can be extended to generalized linear models, although we cannot take advantage of the explicit-solution shortcuts without making approximations

One-step approximations

- The resulting approximate statistics are said to be *one-step approximations* to the true values
- The issue is that we can quickly calculate the one-step approximations based on the current weights $\{w_i\}$ without refitting anything, but to calculate the exact value, we would need to go through n IRLS algorithms
- The approximations are usually pretty good, although if one point has a very large influence, then the approximation may be quite different from the true value

One-step approximations

One-step approximations allow us to quickly calculate the following diagnostic statistics for GLMs:

- Studentized deleted residuals
- Δ_β (for assessing the change in individual coefficients)
- Cook's distance (for assessing overall influence over the model fit)

Variance inflation factors

- It is worth mentioning variance inflation factors (VIF) briefly here
- VIF is a function of $\mathbf{X}$ alone, and therefore how VIF is calculated and what it means is essentially equivalent to the linear regression case ("essentially equivalent" because we do have weights for GLMs)
- In R, we can use the `vif` function from the `car` package:

```
> vif(fit)
      Age       Sex   Age:Sex
 7.416253 14.159377 16.989516
```

- In SAS, this is a bit painful, as we have to use `PROC REG`, which doesn't support the `CLASS` statement or interactions in the `MODEL` statement, and you have to calculate and incorporate the weights manually (see code for the messy details)

Multicollinearity

- If you believe multicollinearity to be a problem, it is often a good idea to look at the correlation matrix for $\mathbf{X}$:

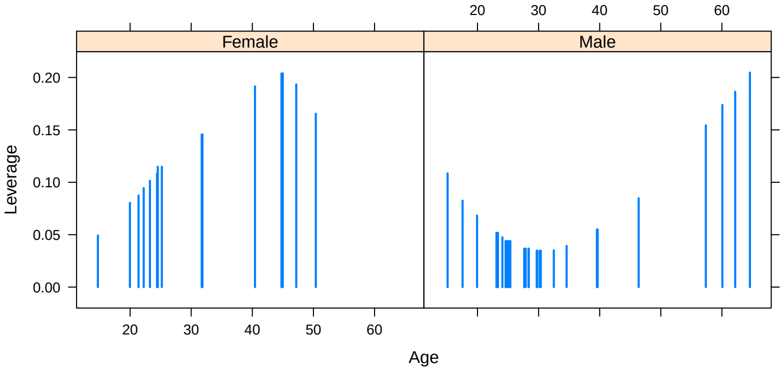
```
cor(model.matrix(fit)[,-1])
```

	Age	Sex	Age:Sex
Age	1.00	0.04	0.52
Sex	0.04	1.00	0.82
Age:Sex	0.52	0.82	1.00

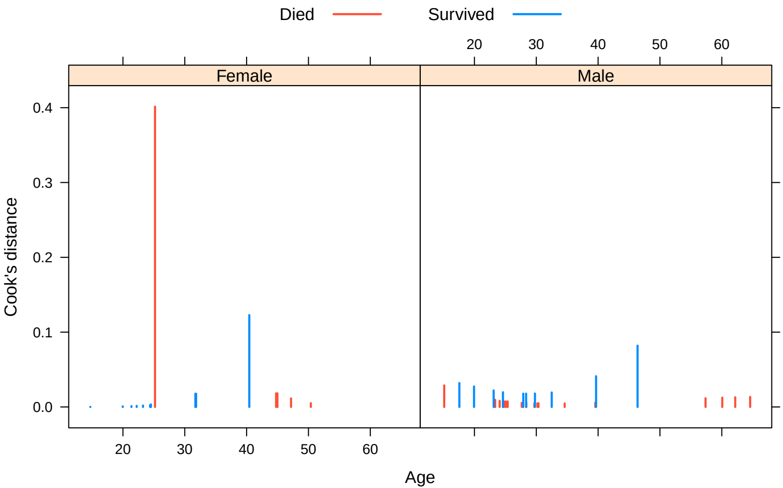
- In this model, we are certainly introducing a lot of variability by including an interaction; on the other hand, the interaction did seem to be important $p = 0.05$

Leverage

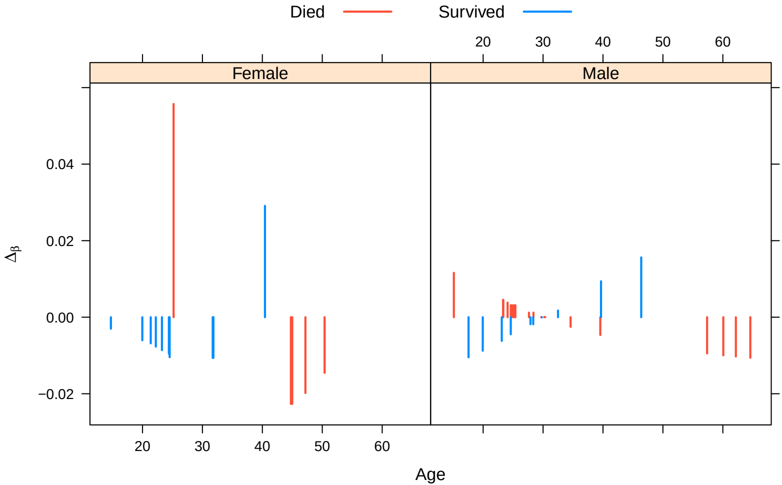
To get a sense of the information these statistics convey, let's look at various plots of the Donner party data, starting with leverage:



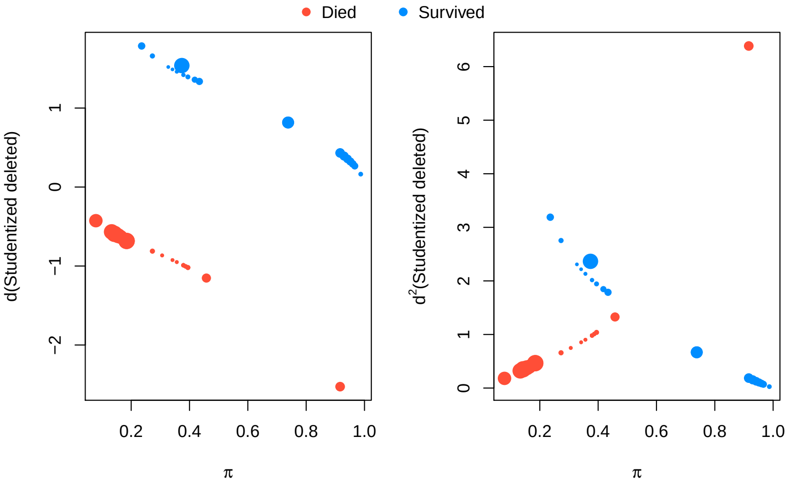
Cook's Distance



Delta-beta (for effect of age)



Residuals / proportional leverage



Summary

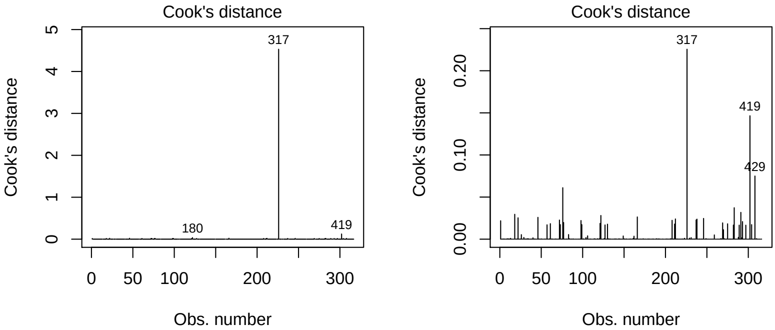
- Residuals are certainly less informative for logistic regression than they are for linear regression: not only do yes/no outcomes inherently contain less information than continuous ones, but the fact that the adjusted response depends on the fit hampers our ability to use residuals as external checks on the model
- This is mitigated to some extent, however, by the fact that we are also making fewer distributional assumptions in logistic regression, so there is no need to inspect residuals for, say, skewness or heteroskedasticity

Summary (cont'd)

- Nevertheless, issues of outliers and influential observations are just as relevant for logistic regression as they are for linear regression
- In my opinion, it is almost never a waste of time to inspect a plot of Cook's distance
- If influential observations are present, it may or may not be appropriate to change the model, but you should at least understand why some observations are so influential

Extubation example

Left: Linear cost; Right: Log(Cost)



Variance inflation

- Finally, keep in mind that although multicollinearity and variance inflation are important concepts, it is not always necessary to calculate a VIF to assess them
- It is usually a good idea when modeling to start with simple models and gradually add in complexity
- If you add a variable or interaction and the standard errors increase dramatically, this is a direct observation of the phenomenon that VIFs are attempting to estimate