Assessing the fit of the binomial regression model

March 3, 2022

1 Assessing the fit of the binomial regression model

In this lesson, we will discuss the deviance as a measure of the goodness of fit. It is important to note that these goodness of fit metrics only hold when the number of trials for our binomial regression is relatively large (say, roughly greater than 5). **In particular, that means that these metrics are not useful for the case where the response is a 0-1 Bernoulli.**

Reminder: Goodness of fit in the normal linear regression framework Recall that, under the assumption that the normal linear model was correct, the coefficient of determination was used as a measure of how well the model fits the data. The coefficient of determination was defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

In this formula, the extent to which the model fits the data is given in the residual sum of squares,

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

After all, the RSS is what we minimized to estimate our parameters, $\boldsymbol{\beta}$, so the smaller it is, the better fit we have (again, assuming that the regression assumptions are correct). The limiting case is that the RSS is zero, which means that every deviation $y_i - \widehat{y}_i = 0$, which means that the data perfectly fall on our regression line/surface, and the model fits perfectly. **In this lesson, our goal will be to find analogous goodness of fit metrics for the binomial regression model. We will also briefly mention analogs to the F-tests from normal linear regression.**

Deviance as a measure of goodness of fit Generally, the *deviance* of a GLM is -2 times the log likelihood of the GLM evaluated at the MLEs:

$$D = -2\ell(\,\widehat{\boldsymbol{\beta}}\,) = -2\sum_{i=1}^{n}\left[y_i\eta_i - n_i\log\left(1+e^{\eta_i}\right) + \log\binom{n_i}{y_i}\right]$$

If our modeling assumptions are correct, a smaller deviance means a better fit, in just the same way that a smaller residual sum of squares meant a better fit in normal regression.

There are a few special cases of the deviance that will be useful for assessing goodness of fit:

1. *The null deviance.* The null deviance is the deviance for the null model - i.e., the model with just an intercept term, and no predictors. In this case, $p_i = \bar{y}$, and the *null deviance* is

$$\begin{aligned} D_{null} &= -2\sum_{i=1}^{n}\left[y_i\eta_i - n_i\log\left(1+e^{\eta_i}\right) + \log\binom{n_i}{y_i}\right] \\ &= -2\sum_{i=1}^{n}\left[y_i\log\left(\frac{\bar{y}}{1-\bar{y}}\right) - n_i\log\left(1+\exp\left\{\log\left(\frac{\bar{y}}{1-\bar{y}}\right)\right\}\right) + \log\binom{n_i}{y_i}\right] \\ &= -2\sum_{i=1}^{n}\left[y_i\log\left(\frac{\bar{y}}{1-\bar{y}}\right) - n_i\log\left(1+\left(\frac{\bar{y}}{1-\bar{y}}\right)\right) + \log\binom{n_i}{y_i}\right] \end{aligned}$$

2. *The saturated deviance.* This is the deviance of the saturated model - i.e., the model where each data point has it's own unique parameter. In this case, the MLE is $\widehat{p}_i = y_i/n_i$.

$$D_{sat} = -2\sum_{i=1}^{n}\left[y_i\eta_i - n_i\log\left(1+e^{\eta_i}\right) + \log\binom{n_i}{y_i}\right] = -2\sum_{i=1}^{n}\left[\right]$$

3. *The residual deviance.* The residual deviance is the difference between the deviance for a given model of interest - e.g., the one you've fit for your data in R - and the saturated model. Let's use the notation D_p for the deviance of the model of interest (one with p predictors). Then the residual deviance can be shown to be:

$$D_{resid} = D_p - D_{sat} = -2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)\right]$$

Importantly, when $n_i > 5$ for all $i = 1, ..., n$, it can be shown that the residual deviance (and more generally, any difference in deviances) follows a χ^2 distribution. Under the null hypothesis that the p model fits the data, the degrees of freedom will be the degrees of freedom of D_p minus the degrees of freedom of D_{sat}. So, in our case: $df(D_{resid}) = (n - (p+1)) - (n - n) = n - (p+1)$. Thus, we can use $D_{resid} \sim \chi^2\big(n - p + 1\big)$ in a test of the fit of our model. The hypotheses under consideration are:

H_0 : The model with p parameters fits well enough. *vs.* H_1 : The model with p parameters does not fit well en

We will reject the null hypothesis when D_{resid} is too large (an upper-tailed chi-squared test).

1.0.1 Example: Challenger data

The 1986 crash of the space shuttle Challenger was linked to failure of O-ring seals in the rocket engines. Data was collected on the 23 previous shuttle missions. The launch temperature on the day of the crash was 31F. The `orings` dataframe contains 23 observations on the following 2 variables.

1. `temp`: temperature at launch in degrees F
2. `damage`: number of damaged o-ring seals out of $n_i = 6$ possible seals, for all $i = 1, ..., 23$.

Let's perform a binomial regression (with the logit link function) using these data, and perform the deviance test described above. First, we'll read in the data:

[71]:
```
library(ggplot2)
# Load the data
orings = read.csv("orings.txt", sep="")
summary(orings)
length(orings$damage)
```

```
      temp           damage
 Min.   :53.00   Min.   :0.0000
 1st Qu.:67.00   1st Qu.:0.0000
 Median :70.00   Median :0.0000
 Mean   :69.57   Mean   :0.4783
 3rd Qu.:75.00   3rd Qu.:1.0000
 Max.   :81.00   Max.   :5.0000
```

23

Now, let's run the a binomial regression. Note that we specify the response as a two-column matrix with the columns giving the numbers of successes and failures.

[60]:
```
n = 6;
glmod = glm(cbind(orings$damage, n-orings$damage) ~ temp, data = orings, family␣
 ↪= binomial)
summary(glmod)
```

```
Call:
glm(formula = cbind(orings$damage, n - orings$damage) ~ temp,
    family = binomial, data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9529  -0.7345  -0.4393  -0.2079   1.9565

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
```

```
Residual deviance: 16.912  on 21  degrees of freedom
AIC: 33.675

Number of Fisher Scoring iterations: 6
```

Under the null hypothesis that this model fits the data well enough, $D_{resid} \sim \chi^2(n-p+1) = \chi^2(23-2) = \chi^2(21)$ in a test of the fit of our model. The hypotheses under consideration are:

H_0 : The model with p parameters fits well enough. *vs.* H_1 : The model with p parameters does not fit well en

We can use the `pchisq()` function to calculate the p-value for this test. We may pull the residual deviance and degrees of freedom from the table above, or use some built in functions to extract them:

[73]:
```
pchisq(16.912, 21, lower = FALSE) #pull the values from the table above
#pchisq(deviance(glmod), df.residual(glmod), lower = FALSE) # use built in␣
↪functions to calculate the deviance and df
```

0.71642667191193

Here, we see that the p-value is large, and so we do not have evidence against the null hypothesis. Thus, we might conclude that the current model with temperature as a predictor fits the data well. We note that as temperature increases by one degree F, the odds of failure are changed by a factor of $e^{-0.21623} = 0.80555$. That is, if the odds of failure at temperature t are o, then if we increase the temperature by one degree F to $t+1$, our odds of failure are about $0.805o$, or roughly 80.5% of what they were. Of course, if we want to draw a causal conclusion, we'd need to impose further assumptions on the data!

[ ]:
```

```