Lecture 15- Deviance and Model Selection

Peter Shaffery

3/4/2021

Wald Testing

Two lectures ago we looked at the sampling distribution for $\hat{\beta}$, and showed that it was asymptotically normal, with mean $\vec{\beta}^*$ (the true parameter vector). Based on this we could, in principal, calculate the standard error s.e.$(\hat{\beta})$.

For the hypothesis $H_0 : \beta_j = 0$ we can calculate the test statistic:

$$z_j = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)}$$

Allowing us to compute a p-value, testing whether or not a single variable "belongs" in a model, given every other coefficient in the model.

This is known as a **Wald test**, and by default R computes it in the `glm` regression table (as we have seen).

But what if we want to determine whether to include *multiple* variables in a model. For linear regression we could use tools like Mallow's C_p or the adjusted R^2 to determine whether the added variables were "worth it", from a predictive perspective. One thing we didn't mention then, however (which did accidently pop up in Hwk 2…) is that we can *also* formulate this problem as a hypothesis test.

Deviance

Last lecture we saw a few ways that we could determine model fit, including McFadden's R^2:

$$R^1 = 1 - \frac{\log L}{\log L_0}$$

This worked okay, but it was unclear what a "good" value of this R^2 was. Obviously "large" is good, but how large?

A conceptually similar quantity, which will allow us to answer this question, is the **likelihood ratio statistic**

$$\lambda = \frac{L(\vec{\beta}_{\text{max}})}{L(\hat{\beta})}$$

Here we are comparing the likelihood of the MLE model $L(\hat{\beta})$ to the likelihood of a *saturated* model $L(\vec{\beta}_{\text{max}})$, rather than a null model L_0.

The saturated model is "largest" possible model with the same link function and distribution as the MLE model. If we have n observations y_i, then the largest possible model assigns a parameter to each y_i, ie. it is the model with n parameters. It is not possible to fit a model with more than n parameters (the problem will be *underconstrained*).

Rather than working with λ directly, it is much more common to deal with its logarithm:

$$\log \lambda = l(\beta_{\text{max}}) - l(\hat{\beta})$$

When $\log \lambda$ is large and positive it indicates that $\hat{\beta}$ is doing a relatively poor job fitting the data, relative to a more complex model.

A really nice fact about $\log \lambda$ is that we can actually derive its sampling distribution (after a little scaling), using our knowledge of the sampling distribution of $\hat{\beta}$. When n is large:

$$2 \log \lambda \to \chi^2(p - m, v)$$

Where p is the number of parameters in β_{max} (often $p = n$) and m is the number of parameters in $\hat{\beta}$. v is a *non-centrality* parameter which will be 0 if $\hat{\beta}$ is almost as good as β_{max}.

Knowing the sampling distribution $2 \log \lambda$ will be useful for performing hypothesis testing with basically any GLM (as will see shortly). Because of this, we give $D = 2 \log \lambda$ the special name of *Deviance*.

Derivation of the Deviance Sampling Distribution

Let $\vec{\beta}^*$ and $\vec{\beta}^*_{\text{max}}$ denote the *true* parameter values of $\hat{\beta}$ and β_{max} respectively. For each of these quantities, observe that we can use a Taylor approximation to write:

$$2\left(l(\hat{\beta}) - l(\vec{\beta}^*)\right) \approx (\vec{\beta}^* - \hat{\beta})^T J(\hat{\beta})(\vec{\beta}^* - \hat{\beta})$$
$$2\left(l(\beta_{\text{max}}) - l(\vec{\beta}^*_{\text{max}})\right) \approx (\vec{\beta}^*_{\text{max}} - \beta_{\text{max}})^T J(\beta_{\text{max}})(\vec{\beta}^*_{\text{max}} - \beta_{\text{max}})$$

Here $J(\hat{\beta})$ is a matrix of second derivatives:

$$J(\hat{\beta}) = \begin{bmatrix} \frac{d^2 l}{d\beta_1^2} & \cdots & \frac{d^2 l}{d\beta_1 d\beta_m} \\ \vdots & \ddots & \vdots \\ \frac{d^2 l}{d\beta_m d\beta_1} & \cdots & \frac{d^2 l}{d\beta_m^2} \end{bmatrix}$$

Which plays the same role in the variance of the vector valued $\hat{\beta}$ as the scalar second derivative did for β_1.

Since $\hat{\beta}$ is asymptotically normal, it can be shown that:

$$2\left(l(\hat{\beta}) - l(\vec{\beta}^*)\right) \approx (\vec{\beta}^* - \hat{\beta})^T J(\hat{\beta})(\vec{\beta}^* - \hat{\beta}) \to \chi^2(m)$$
$$2\left(l(\beta_{\text{max}}) - l(\vec{\beta}^*_{\text{max}})\right) \approx (\vec{\beta}^*_{\text{max}} - \beta_{\text{max}})^T J(\beta_{\text{max}})(\vec{\beta}^*_{\text{max}} - \beta_{\text{max}}) \to \chi^2(p)$$

Now, this is not exactly the form of the deviance, but it's close. A little algebra gives us:

$$\begin{aligned} D &= 2\left(l(\beta_{\text{max}}) - l(\hat{\beta})\right) \\ &= 2\left(l(\beta_{\text{max}}) - l(\vec{\beta}^*_{\text{max}})\right) - 2\left(l(\hat{\beta}) - l(\vec{\beta}^*)\right) + 2\left(l(\beta^*_{\text{max}}) - l(\vec{\beta}^*)\right) \end{aligned}$$

A neat property of χ^2 random variables is, if:

$$X \sim \chi^2(a)$$
$$Y \sim \chi^2(b)$$

With $a > b$, then $X - Y \sim \chi^2(a - b)$. Moreover, if we introduce a constant c, then $X + c \sim \chi^2(a, c)$, which we refer to as a **non-central* χ^2 distribution.

We therefore have that:

$$D \to \chi^2(m - p, v)$$

Where the centrality parameter $v = 2\left(l(\beta^*_{\text{max}}) - l(\vec{\beta}^*)\right)$. Hence, when $\hat{\beta}$ and β_{max} are nearly equal in performance $v \approx 0$.

Hypothesis Testing with Deviance

Say that we have two parameter vectors $\vec{\beta}_0 = [\beta_1, ..., \beta_k]^T$ and $\vec{\beta}_1 = [\beta_1, ..., \beta_m]^T$. We see that the models represented by these vectors are **nested** if every element of β_0 is an element of β_1. In this case, we can think of $\vec{\beta}_0$ as a *sub-model* of $\vec{\beta}_1$. Denote the corresponding models by M_0 and M_1.

Deviance hypothesis testing allows us to pick between M_0 and M_1:

1. H_0 : model M_0 is "correct" (ie. parameters $\beta_{k+1} = ... = \beta_m = 0$)
2. H_A : model M_1 is correct (ie. at least one of the parameters $\beta_{k+1}, ..., \beta_m$ are not equal to 0)

The way we can compare between these hypotheses is through the *difference* in model deviances. Let D_i correspond to the model deviace of M_i, and then define:

$$\begin{aligned} \Delta D &= D_0 - D_1 \\ &= 2\left(l(\hat{\beta}_1) - l(\hat{\beta}_0)\right) \end{aligned}$$

Now, there are two cases that we care about:

Case 1- Both models perform equally well

If both models perform equally well relative to $\beta_{\max}$, then both v_0 and v_1 will be approximately 0, and hence:

$$\begin{aligned} D_0 &\sim \chi^2(n-k) \\ D_1 &\sim \chi^2(n-m) \\ \Delta D &\sim \chi^2(m-k) \end{aligned}$$

Now, if $\vec{\beta}_0$ and $\vec{\beta}_1$ are equally performant, we should prefer $\vec{\beta}_0$ since it is the simpler model (this is sometimes called the *principal of parsimony*). Hence if the actual value of ΔD is consistent with the distribution $\chi^2(m-k)$ then we fail to reject H_0.

Case 2- M_0 is worse than M_1

If M_0 performs *worse* than $\beta_{\max}$ then v_0 will be greater than 0. This will result in a value of ΔD that is **larger** than would be expected under $\chi^2(m-k)$. Hence if ΔD lies in the upper tails of $\chi^2(m-k)$ then we reject H_0 in favor of H_1.

Putting it All Together

We therefore have the following procedure for performing a hypothesis between M_0 and M_1:

1. Choose a significance level α
2. Compute ΔD
3. If $Pr[X \geq \Delta D] \leq \alpha$ for $X \sim \chi^2(m-k)$ then reject H_0. Otherwise fail to reject H_0.

Example: Titanic dataset

The Titanic dataset contains individual outcomes of about half the passangers aboard the famous Titanic- a passanger ship which sank in 1912. The dataset contains 1132 records, and includes a number of variables about each passenger as well as whether they survived or not:

```r
library(tidyverse)
library(magrittr)

dat = read.csv('../../data/titanic.csv')

# the titanic data contains some missing data that we'll just ignore for now....
```

```r
dat %<>% drop_na

# we don't care about a few of the columns
dat %<>% select(-c('name','ticket','cabin'))

dat %>% head
```

```
##   survived pclass    sex     age sibsp parch     fare embarked
## 1        1      1 female 29.0000     0     0 211.3375        S
## 2        1      1   male  0.9167     1     2 151.5500        S
## 3        0      1 female  2.0000     1     2 151.5500        S
## 4        0      1   male 30.0000     1     2 151.5500        S
## 5        0      1 female 25.0000     1     2 151.5500        S
## 6        1      1   male 48.0000     0     0  26.5500        S
```

```r
dat$pclass %<>% as.factor
```

Let's use a deviance hypothesis test at $\alpha = .95$ to compare two models:

1. M_0: a model which includes PCLASS, SEX, and AGE
2. M_1: a model which includes PCLASS, SEX, AGE, SIBSP, and PARCH

```r
k = 3
m = 5
alpha = .95

mod0 = glm(survived~pclass+age+sex, family=binomial, data=dat)
mod1 = glm(survived~pclass+age+sex+sibsp+parch, family=binomial, data=dat)

D0 = mod0$deviance
D1 = mod1$deviance

delta.D = D0 - D1
p = pchisq(delta.D, m-k, lower.tail=FALSE)
print(p)
```

```
## [1] 0.002100833
```

Where therefore reject M_0 in favor of M_1!

Importantly, be aware that this does not mean that **every** variable in M_1 is important, look at the summary table:

```r
summary(mod1)
```

```
##
## Call:
## glm(formula = survived ~ pclass + age + sex + sibsp + parch,
##     family = binomial, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7009  -0.6644  -0.4202   0.6663   2.5192
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.904220   0.362085  10.783  < 2e-16 ***
## pclass2     -1.366033   0.230039  -5.938 2.88e-09 ***
## pclass3     -2.350481   0.228899 -10.269  < 2e-16 ***
## age         -0.039436   0.006639  -5.940 2.85e-09 ***
## sexmale     -2.556308   0.173269 -14.753  < 2e-16 ***
## sibsp       -0.352838   0.105340  -3.350  0.00081 ***
## parch        0.074320   0.099898   0.744  0.45690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1413.57  on 1044  degrees of freedom
## Residual deviance:  970.05  on 1038  degrees of freedom
## AIC: 984.05
##
## Number of Fisher Scoring iterations: 4
```

Based on this it would appear that just SIBSP is significant.

Deviance Testing for Linear Regression

The calculations that we performed to get the distribution of ΔD hold for any GLM. However, depending on your choice of model you may not be able to compute $l(\hat{beta})$ directly.

Binomial Deviance

For a binomial model we can see that computing D is straightforward.

Say that we have observations $(\vec{x}_i, y_i, n_i)$ where y_i is the number of "successes" in n_i trials. Let $\hat{\pi}_i = \text{logit}^{-1}(\vec{x}_i^T \hat{\beta})$. We therefore have that:

$$l(\hat{\beta}) = \sum_i y_i \log \hat{\pi}_i + (n_i - y_i) \log (1 - \hat{\pi}_i) + \log \binom{n_i}{y_i}$$

Now, for binomial regression that saturated model is the model where we estimate each π_i with the observed success rate $\frac{y_i}{n_i}$, hence:

$$l(\beta_{\max}) \sum_i y_i \log \frac{y_i}{n_i} + (n_i - y_i) \log (1 - \frac{y_i}{n_i}) + \log \binom{n_i}{y_i}$$

Once we have computed the MLE $\hat{\beta}$, we can compute $D = l(\beta_{\max}) - l(\hat{\beta})$, which simplifies to:

$$D = 2 \sum \left[y_i \log \frac{y_i}{n_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i(1 - \hat{\pi}_i)} \right]$$

Notice that for this model the Deviance can be calculated using only quantities which we have directly observed. There is nothing else we need to estimate, beyond $\hat{\beta}$.

Normal Deviance

This is not the case for the normal distribution. Say that we have computed an MLE $\hat{\beta}$, and would now like that compute $l(\hat{\beta})$:

You may recall that the normal log-likelihood for the MLE is:

$$l(\hat{\beta}) = \frac{-1}{2\sigma^2} \sum (y_i - \vec{x}_i^T \hat{\beta})^2 - \frac{1}{2} N \log(2\pi\sigma^2)$$

(We originally just ignored the $\frac{1}{2}N\log(2\pi\sigma^2)$ since it didn't involved $\vec{\beta}$, but here it will make a difference.)

Now, the saturated model for linear regression assigns each y_i it's own mean value μ_i. It's not hard to convince yourself that the "best-fitting" saturated model would therefore just set $\mu_i = y_i$, which gives us:

$$l(\beta_{\text{max}}) = \frac{1}{2}N\log(2\pi\sigma^2)$$

The deviance for the normal model is therefore:

$$D = \frac{-1}{2\sigma^2}\sum(y_i - \vec{x_i}^T\hat{\beta})^2$$

Unfortunately, we cannot compute this directly, since it depends on σ^2. Instead, we need to *estimate* the deviance using an estimate for σ^2. The usual estimator is (as we have seen):

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\epsilon}_i^2}{N-m}$$

Where $\hat{\epsilon}_i$ are the model residuals.

For a normal model, D itself is actually *exactly* χ^2 distributed. However, the estimator:

$$\hat{D} = \frac{-1}{2\hat{\sigma}^2}\sum(y_i - \vec{x_i}^T\hat{\beta})^2$$

Is not χ^2. Instead, since both the numerator and the denominator are χ^2, $\hat{D}$ actually has an F-distribution. Therefore, deviance hypothesis testing for linear regression takes the form of an *F-test* (see RABE 3.10.2, or IGLM Ch. 6).

Akaike Information Criterion

Now, it is often the case that we want to compare between *non-nested* models. We can no longer use a deviance test in this case. Instead, we turn to the Akaike Inforamtion Criterion (AIC). This quantity is defined:

$$\text{AIC} = -2l(\hat{\beta}) + 2p$$

Note the relationship to deviance:

$$AIC = D - 2l(\beta_{\text{max}}) + 2p$$

If we're comparing multiple models using the AIC, then we really only care about $\Delta\text{AIC} = \text{AIC}_1 - \text{AIC}_2$. For both models the $-2l(\beta_{\text{max}})$ term will be the same, so we see that:

$$\Delta\text{AIC} = (D_1 - D_2) + 2(p_1 - p_2) = \Delta D + 2\Delta p$$

So using AIC to pick between models is fundamentally the same "kind" of comparison as Deviance, except we're penalizing for the difference in model complexity Δp.