Ve406 Lecture 18

Jing Liu

UM-SJTU Joint Institute

July 18, 2018

- Logistic regression can be understood as a particular case of

  Generalised linear Model (GLM)

- In a linear model, we model the conditional mean

$$\mathbb{E}\left[Y \mid \mathbf{X}\right]$$

  by a linear function

$$\mathbb{E}\left[Y \mid \mathbf{X}\right] = m\left(\mathbf{X}\right)$$

- In its simplest form,

$$m\left(\mathbf{X}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \mathbf{X}\boldsymbol{\beta}$$

  or more generally

$$m\left(\mathbf{X}\right) = \beta_0 + \beta_1 f(x_1) + \beta_2 f(x_2) + \cdots + \beta_k f(x_k) = \mathbf{X}^{*}\boldsymbol{\beta}$$

- Of course, not all conditional mean can be well modelled by a linear function

$$m\left(\mathbf{X}\right)=\mathbf{X}\boldsymbol{\beta}\qquad\text{or}\qquad m\left(\mathbf{X}\right)=\mathbf{X}^{*}\boldsymbol{\beta}$$

- The idea of GLM is to consider modelling a transformation of

$$\mu=\mathbb{E}\left[Y\mid\mathbf{X}\right]$$

  by a linear function, that is,

$$g\left(\mu\right)=m\left(\mathbf{X}\right)\iff\mu=g^{-1}\left(m\right)$$

  where g is known as the link function of the generalised linear model.

- In logistic regression, the link is the logit function

$$\mu=\mathbb{E}\left[Y\mid\mathbf{X}\right]=\Pr\left[Y=1\mid\mathbf{X}\right];\qquad g(\mu)=\ln\left(\frac{\mu}{1-\mu}\right)$$

$$\implies\mu=\frac{\exp(m)}{1+\exp(m)}=\frac{\exp\left(\mathbf{X}\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}\boldsymbol{\beta}\right)}=\frac{\exp\left(\mathbf{X}^{*}\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}^{*}\boldsymbol{\beta}\right)}$$

- In terms of GLM, logistic regression is also known as Binomial regression

$$Y_i \mid \mathbf{X} \sim \text{Binomial}\left(n_i, g^{-1}(m)\right)$$

- Clearly, if the identify function is assumed,

$$g(\mu) = \mu$$

  we just have the ordinary linear regression

$$Y_i \mid \mathbf{X} \sim \text{Normal}\left(g^{-1}(m), \sigma^2\right)$$

- Instead of being given a model, suppose we try to come up with a model for

$$Y_i \mid \mathbf{X} \sim \text{Poisson}\left(\lambda\right)$$

Q: What should we use for the link function in this case?

- Recall the density function of a Poisson random variable is given by

$$\Pr\left[Y_i = y_i\right] = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \qquad \text{and} \qquad \mu = \mathbb{E}\left[Y_i\right] = \lambda > 0$$

- If we use the natural logarithmic function as the link function, then

$$\ln\left(\mathbb{E}\left[Y \mid \mathbf{X}\right]\right) = m\left(\mathbf{X}\right)$$
$$\implies \mu = \lambda = \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)$$

which ensures $\mu = \lambda$ is always positive, and parameters

$$\begin{aligned}\mu &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2(x_2 + 1) + \cdots + \beta_k x_k\right)\\ &= \exp(\beta_2)\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)\end{aligned}$$

can be interpreted in a reasonably simple fashion with a reasonably simple log-likelihood function to maximise. This is known as Poisson regression.

- To illustration Poisson regression, consider the following dataset:

COUNT	number of accidents
INB	inner burden thickness
EXTRP	percentage of coal extracted from the mine
AHS	the average height of the coal seam in the mine
AGE	the age of the mine

- The idea is to use Poisson regression to model the number of accidents to understand whether the variables collected from 44 coal mines in USA in a 3 month period can explain the number of accidents in those mines.

```
> str(mines.df)
```

```
'data.frame':   44 obs. of  5 variables:
 $ COUNT: int  2 1 0 4 1 2 0 0 4 4 ...
 $ INB  : int  50 230 125 75 70 65 65 350 350 160 ...
 $ EXTRP: int  70 65 70 65 65 70 60 60 90 80 ...
 $ AHS  : int  52 42 45 68 53 46 62 54 54 38 ...
 $ AGE  : num  1 6 1 0.5 0.5 3 1 0.5 0.5 0 ...
```

```
> mines.pois =
+   glm(COUNT ~ INB + EXTRP + AHS + AGE,
+       family=poisson, data=mines.df)

> summary(mines.pois)
```

```
Call:
glm(formula = COUNT ~ INB + EXTRP + AHS + AGE, family = poisson,
    data = mines.df)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept) -3.6097078  1.0284740  -3.510 0.000448 ***
INB         -0.0014441  0.0008415  -1.716 0.086145 .
EXTRP        0.0622011  0.0122872   5.062 4.14e-07 ***
AHS         -0.0017578  0.0050737  -0.346 0.729003
AGE         -0.0296244  0.0163143  -1.816 0.069394 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.984  on 43  degrees of freedom
Residual deviance: 37.717  on 39  degrees of freedom
AIC: 143.99

Number of Fisher Scoring iterations: 5
```

- For generalised linear models, it is conventional to report

  deviance

  which is a measure of of how well a model fits the data, and is defined

  $$D = 2\left(\ell_{sat} - \ell_{prop}\right)$$

  where ℓ_{prop} is the log-likelihood of our proposed model at MLE for the given data, and ℓ_{sat} is the log-likelihood of the saturated model, that is, one parameter per data point, which will have the largest possible log-likelihood.

- If the proposed model fits the data reasonably well, then

  $$D \sim \chi^2_{n-(k+1)}$$

  where n is the number of observations, $k+1$ is the number of parameters in the proposed model, the factor of 2 in the definition is to ensure this.

Q: Can you see that it gives a way to check the goodness of fit for our model?

```
> summary(mines.pois)
```

```
Call:
glm(formula = COUNT ~ INB + EXTRP + AHS + AGE, family = poisson,
    data = mines.df)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept) -3.6097078  1.0284740  -3.510 0.000448 ***
INB         -0.0014441  0.0008415  -1.716 0.086145 .
EXTRP        0.0622011  0.0122872   5.062 4.14e-07 ***
AHS         -0.0017578  0.0050737  -0.346 0.729003
AGE         -0.0296244  0.0163143  -1.816 0.069394 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.984  on 43  degrees of freedom
Residual deviance: 37.717  on 39  degrees of freedom
AIC: 143.99

Number of Fisher Scoring iterations: 5
```

- The deviance of our model is given under the name “Residual deviance”

```
> 1-pchisq(37.717, 39)
```

```
[1] 0.5283455
```

- Since this is applicable to any GLM, it works for logistic regression as well,

```
> summary(credit.all3.LG)
```

```
Call:
glm(formula = default ~ balance + income + student, family = binomial,
    data = credit.df)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

```
> 1-pchisq(1571.5, 9996)
```

```
[1] 1
```

- For a similar reason, the concept of deviance can be used to compare GLMs

$$\begin{aligned} D_{sub} - D_{prop} &= 2\left(\ell_{sat} - \ell_{sub}\right) - 2\left(\ell_{sat} - \ell_{prop}\right) \\ &= 2\left(\ell_{prop} - \ell_{sub}\right) \end{aligned}$$

- This difference represents the increase in the deviance if d number of terms are dropped from the original proposed model or the full model.

$$\left(D_{sub} - D_{prop}\right) \sim \chi_d^3$$

- Recall we had the following result

```
> summary(credit.all3.LG)
```

```
Call:
glm(formula = default ~ balance + income + student, family = binomial,
    data = credit.df)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

- Hence consider the following

```
> credit.sub.LG = glm(
+    default~balance+student, family = binomial,
+    data = credit.df)
```

```
> dd = credit.sub.LG$deviance -
+    credit.all3.LG$deviance
```

- While this difference will always be positive, if the increase is small then dropping the extra terms will not increase the deviance by very much and so the variables can be dropped in the interests of getting a simpler model.

```
> 1-pchisq(dd, 1)
```

```
[1] 0.7115139
```

- The underlying assumption of the χ^2 test is that the sub model is adequate as well in comparison to the original proposed model, so can drop `income`.

- Now for the Poisson regression model we had,

```
> summary(mines.pois)
```

```
Call:
glm(formula = COUNT ~ INB + EXTRP + AHS + AGE, family = poisson,
    data = mines.df)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept) -3.6097078  1.0284740  -3.510 0.000448 ***
INB         -0.0014441  0.0008415  -1.716 0.086145 .
EXTRP        0.0622011  0.0122872   5.062 4.14e-07 ***
AHS         -0.0017578  0.0050737  -0.346 0.729003
AGE         -0.0296244  0.0163143  -1.816 0.069394 .
```

```
> mines.sub.pois =
+    glm(COUNT ~ EXTRP + AGE,
+        family=poisson, data=mines.df)
>
> dd = mines.sub.pois$deviance - mines.pois$deviance
>
> 1-pchisq(dd, 2)
```

```
[1] 0.1416521
```

```
> summary(mines.sub.pois)
```

```
Call:
glm(formula = COUNT ~ EXTRP + AGE, family = poisson, data = mines.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8023  -0.8554  -0.1627   0.4437   2.4878

Coefficients:
            Estimate Std. Error z value Pr(>z)
(Intercept) -3.58943    0.94471  -3.800 0.000145 ***
EXTRP        0.05875    0.01169   5.027 4.98e-07 ***
AGE         -0.03802    0.01545  -2.460 0.013888 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.984  on 43  degrees of freedom
Residual deviance: 41.626  on 41  degrees of freedom
AIC: 143.9

Number of Fisher Scoring iterations: 5
```

- Notice that R gives the link function $g\left(\mathbb{E}\left[Y \mid \mathbf{X}\right]\right)$ for prediction by default.

```
> predict(mines.sub.pois, data.frame(
+   EXTRP = c(50, 90), AGE = c(2, 3)),
+   type = "response")
```

```
        1         2
0.4828544 4.8738560
```

```
> predict(mines.sub.pois, data.frame(
+   EXTRP = c(50, 90), AGE = c(2, 3)))
```

```
         1          2
-0.7280402  1.5838854
```

```
> predict(credit.sub.LG, data.frame(
+   balance = c(1000, 2000), student = "Yes"),
+         type = "response")
```

```
          1           2
0.003248626 0.502958678
```

```
> predict(credit.sub.LG, data.frame(
+   balance = c(1000, 2000), student = "Yes"))
```

```
          1           2
-5.72626932  0.01183485
```