Logistic Regression Assignment Solutions

David M. Rocke

May 30, 2017

Suppose we have data on 100 cases of myocardial infarction and 150 healthy individuals (`mi = 1 if MI, 0 otherwise`) matched to the MI group by age and sex. From their medical records before the MI (if they had one), we classify the individuals as diabetic, metabolic disorder, and normal blood glucose (`bg = norm, metdis, diabetic`). The table on the next page shows the number of individuals in each group.

	norm	metdis	diabetic	Total
Control	85	50	15	150
MI	35	30	35	100
Total	120	80	50	250

- Find the odds ratio for MI for diabetic individuals vs. normal individuals (ignoring the metabolic disorder individuals). Interpret.

$$\frac{35/15}{35/85} = \frac{85}{15} = 5.67$$

The odds of an MI are 5.67 times higher in a diabetic individual than an individual with normal blood glucose.

- Write down the logistic regression model formulation in detail for predicting MI from bg. Specifically make sure you have defined the coefficients in the model. Use "normal" as the base level for the bg factor.

$$
\begin{aligned}
p &= \Pr(MI|bg) \\
\ln\left[\frac{p}{1-p}\right] &= \beta_0 + \beta_{\text{metdis}} x_{\text{metdis}} + \beta_{\text{diabetic}} x_{\text{diabetic}} \\
x_{\text{metdis}} &= 1 \quad \text{iff bg} = \text{metdis} \\
x_{\text{diabetic}} &= 1 \quad \text{if bg} = \text{diabetic}
\end{aligned}
$$

$$
\begin{aligned}
\beta_{\text{metdis}} &= \text{log-odds ratio of metdis vs. normal} \\
\beta_{\text{diabetic}} &= \text{log-odds ratio of diabetic vs. normal}
\end{aligned}
$$

- Derive the likelihood equation for the model.

$$
\begin{aligned}
\ell^{-1}(x) &= \frac{1}{1+\exp(-x)} \\
p_0 &= \ell^{-1}(\beta_0) \\
p_1 &= \ell^{-1}(\beta_0 + \beta_{\text{metdis}}) \\
p_2 &= \ell^{-1}(\beta_0 + \beta_{\text{diabetic}}) \\
L(\beta_0, \beta_{\text{metdis}}, \beta_{\text{diabetic}}) &= \binom{120}{35} p_0^{35}(1-p_0)^{85} \\
&\quad \times \binom{80}{30} p_1^{30}(1-p_1)^{50} \\
&\quad \times \binom{50}{35} p_2^{35}(1-p_2)^{15}
\end{aligned}
$$

- Derive the maximum likelihood estimates for the parameters of the model, using normal as the default level.

$$
\begin{aligned}
p_0 &= 35/120 = 0.2917 \\
p_1 &= 30/80 = 0.375 \\
p_2 &= 35/50 = 0.70 \\
\beta_0 &= \log[35/85] = -0.8873 \\
\beta_0 + \beta_1 &= \log[30/50] = -0.5108 \\
\beta_0 + \beta_2 &= \log[35/15] = 0.8473 \\
\beta_1 &= -0.5108 - (-0.8873) = 0.3765 \\
\beta_2 &= 0.8473 - (-0.8873) = 1.7346
\end{aligned}
$$

```
> logistic.example
        bg mi non.mi
1     norm 35     85
2   metdis 30     50
3 diabetic 35     15
```

```
> logistic.example.mi
     [,1] [,2]
[1,]   35   85
[2,]   30   50
[3,]   35   15
```

```
> summary(glm(logistic.example.mi~bg,family=binomial,data=logistic.example))

Call:
glm(formula = logistic.example.mi ~ bg, family = binomial, data = logistic.example)

Deviance Residuals:
[1]  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8873     0.2008  -4.418 9.96e-06 ***
bgmetdis      0.3765     0.3061   1.230    0.219
bgdiabetic    1.7346     0.3682   4.711 2.47e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  2.4696e+01  on 2  degrees of freedom
Residual deviance: -1.6653e-14  on 0  degrees of freedom
AIC: 20.031

Number of Fisher Scoring iterations: 3
```

- Compute the maximized log likelihood. What is the deviance? Why?

$$
\begin{aligned}
p_0 &= 35/120 \\
p_1 &= 30/80 \\
p_2 &= 35/50 \\
L &= \binom{120}{35} p_0^{35}(1-p_0)^{85} \\
&\quad \times \binom{80}{30} p_1^{30}(1-p1)^{50} \\
&\quad \times \binom{50}{35} p_2^{35}(1-p_2)^{15}
\end{aligned}
$$

$$
\begin{aligned}
p_0 &= 35/120 \\
p_1 &= 30/80 \\
p_2 &= 35/50 \\
L &= \binom{120}{35} p_0^{35}(1-p_0)^{85} \\
&\quad \times \binom{80}{30} p_1^{30}(1-p1)^{50} \\
&\quad \times \binom{50}{35} p_2^{35}(1-p_2)^{15}
\end{aligned}
$$

```
> p0 <- 35/120
> p1 <- 30/80
> p2 <- 35/50
> ll <- lchoose(120,35)+35*log(p0)+85*log(1-p0)
> ll <- ll + lchoose(80,30)+30*log(p1)+50*log(1-p1)
> ll <- ll + lchoose(50,35)+35*log(p2)+15*log(1-p2)
> print(ll)
[1] -7.015692

> logLik(logistic.example.glm)
'log Lik.' -7.015692 (df=3)
The deviance is zero since this is a saturated model.
```

- If the three parameters are β_0 (the intercept), β_{metdis}, and β_{diabetic} in that order, and if the covariance matrix of the parameters is

$$\begin{pmatrix} 0.04034 & -0.04034 & -0.04034 \\ -0.04034 & 0.09367 & 0.04034 \\ -0.04034 & 0.04034 & 0.13557 \end{pmatrix}$$

  test the hypotheses (separately) that each of the two non-intercept parameters is zero.

$$\beta_0 = -0.8873$$

$$\beta_1 = 0.3765$$

$$\beta_2 = 1.7346$$

$$V = \begin{pmatrix} 0.04034 & -0.04034 & -0.04034 \\ -0.04034 & 0.09367 & 0.04034 \\ -0.04034 & 0.04034 & 0.13557 \end{pmatrix}$$

$$z_1 = 0.3765/\sqrt{0.09367} = 1.230 \quad p = 0.219$$

$$z_2 = 1.7346/\sqrt{0.13557} = 4.711 \quad p = 2.5 \times 10^{-6}$$

```
> summary(glm(logistic.example.mi~bg,family=binomial,data=logistic.example))

Call:
glm(formula = logistic.example.mi ~ bg, family = binomial, data = logistic.example)

Deviance Residuals:
[1]  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8873     0.2008  -4.418 9.96e-06 ***
bgmetdis      0.3765     0.3061   1.230    0.219
bgdiabetic    1.7346     0.3682   4.711 2.47e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  2.4696e+01  on 2  degrees of freedom
Residual deviance: -1.6653e-14  on 0  degrees of freedom
AIC: 20.031

Number of Fisher Scoring iterations: 3
```

- Test the hypothesis that diabetic and metabolic disorder subjects have a log-odds ratio vs. MI of 0.

$$
\begin{aligned}
\beta_0 &= -0.8873 \\
\beta_1 &= 0.3765 \\
\beta_2 &= 1.7346 \\
V &= \begin{pmatrix} 0.04034 & -0.04034 & -0.04034 \\ -0.04034 & 0.09367 & 0.04034 \\ -0.04034 & 0.04034 & 0.13557 \end{pmatrix} \\
x_3 &= (1.7346 - 0.3765) = 1.3581 \\
s_3 &= \sqrt{0.13557 + 0.09367 - 2(0.04034)} = 0.3854 \\
z_3 &= x_3/s_3 = 3.524 \quad p = 4.26 \times 10^{-4}
\end{aligned}
$$

- Find a 95% confidence interval for the odds ratio for MI with respect to normal/diabetic.

$$
\begin{aligned}
\beta_0 &= -0.8873 \\
\beta_1 &= 0.3765 \\
\beta_2 &= 1.7346 \\
V &= \begin{pmatrix} 0.04034 & -0.04034 & -0.04034 \\ -0.04034 & 0.09367 & 0.04034 \\ -0.04034 & 0.04034 & 0.13557 \end{pmatrix} \\
1.7346 &\pm 1.960\sqrt{0.13557} \\
1.7346 &\pm 0.7217 \\
&(1.0129, 2.4563) \quad \text{log-odds ratio} \\
&(2.75, 11.66) \quad \text{odds ratio}
\end{aligned}
$$

- How would you perform the likelihood ratio test for the given model vs. the null model?
- The log-likelihood is -7.015692. We need to compare this to the log-likelihood for the null model, which is where p does not depend on the bg variable. The MLE for p then is $100/250 = 0.40$. Minus twice the difference in these is asymptotically χ^2_2. Calculated using the pooled p, we get a log-likelihood of -19.36386. The test statistic is then $-2[-19.36386 - (-7.015692)] = 24.696$ with $p = 4.34 \times 10^{-6}$.

```
> p <- 100/250

> ll2 <- lchoose(120,35)+35*log(p)+85*log(1-p)
> ll2 <- ll2 + lchoose(80,30)+30*log(p)+50*log(1-p)
> ll2 <- ll2 + lchoose(50,35)+35*log(p)+15*log(1-p)

> print(ll2)
[1] -19.36386

> x2 <- -2*(ll2-ll)
> x2
[1] 24.69634

> pchisq(x2,2,lower=F)
[1] 4.337673e-06

> drop1(logistic.example.glm,test="Chisq")
Single term deletions

Model:
logistic.example.mi ~ bg
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>       0.000 20.031
bg      2   24.696 40.728 24.696 4.338e-06 ***
```