MAST90104: A First Course in Statistical Learning

Week 9 Practical and Workshop Solution

1 Workshop questions

1. Verify that for the binomial regression model with logistic link

$$\begin{aligned}\mathbb{E}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i} &= 0\\ -\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i\partial\theta_j} &= \mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_j}\right)\end{aligned}$$

Solution:

Suppose that $Y_k, k=1,\cdots,n \sim Binomial(m_k, p_k = g^{-1}(\mathbf{x}_k^T\boldsymbol{\theta})$ where $\mathbf{x}_k, i=k,\cdots,n$ are explanatory predictors and $g(p)=\log(\frac{p}{1-p})$ is the logistic link function. Hence

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i} &= \sum_{k=1}^n Y_k\frac{1}{p_k}\frac{\partial p_k}{\partial\theta_i} - (m_k - Y_k)\frac{1}{1-p_k}\frac{\partial p_k}{\partial\theta_i}\\ &= \sum_{k=1}^n \frac{\partial p_k}{\partial\theta_i}Z_k\end{aligned}$$

where $Z_k = Y_k/p_k - (m_k - Y_k)/(1-p_k)$, so $\mathbb{E}(Z_k) = m_kp_k/p_k - m_k(1-p_k)/(1-p_k) = 0$ and var $(Z_k) =$ var $(Y_k/(p_k(1-p_k)) - m_k/(1-p_k)) = m_k/(p_k(1-p_k))$. Thus

$$\begin{aligned}\mathbb{E}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i} &= \sum_{k=1}^n \mathbb{E}\left(\frac{\partial p_k}{\partial\theta_i}Z_k\right)\\ &= \sum_{k=1}^n \frac{\partial p_k}{\partial\theta_i}\mathbb{E}(Z_k) = 0.\end{aligned}$$

Now

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i\partial\theta_j} &= \sum_{k=1}^n \frac{\partial^2 p_k}{\partial\theta_i\partial\theta_j}Z_k + \frac{\partial p_k}{\partial\theta_i}\frac{\partial Z_k}{\partial\theta_j}\\ &= \sum_{k=1}^n \frac{\partial^2 p_k}{\partial\theta_i\partial\theta_j}Z_k + \frac{\partial p_k}{\partial\theta_i}\frac{\partial p_k}{\partial\theta_j}\left(\frac{-Y_k}{p_k^2} - \frac{m_k - Y_k}{(1-p_k)^2}\right)\end{aligned}$$

so

$$\begin{aligned}-\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i\partial\theta_j} &= -0 + \sum_{k=1}^n \frac{\partial p_k}{\partial\theta_i}\frac{\partial p_k}{\partial\theta_j}\left(\frac{m_k}{p_k} + \frac{m_k}{1-p_k}\right)\\ &= \sum_{k=1}^n \frac{\partial p_k}{\partial\theta_i}\frac{\partial p_k}{\partial\theta_j}\frac{m_k}{p_k(1-p_k)}.\end{aligned}$$

Whereas

$$\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_j}\right) = \sum_{k=1}^n \frac{\partial p_k}{\partial\theta_i}Z_k\sum_{l=1}^n\frac{\partial p_l}{\partial\theta_j}Z_l$$

and, given the $Z_k, k = 1 \cdots n$ are independent,

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_j}\right) &= \sum_{k=1}^{n}\frac{\partial p_k}{\partial\theta_i}\mathbb{E}\left(Z_k\sum_{l=1}^{n}\frac{\partial p_l}{\partial\theta_j}Z_l\right)\\
&= \sum_{k=1}^{n}\frac{\partial p_k}{\partial\theta_i}\sum_{l=1}^{n}\frac{\partial p_l}{\partial\theta_j}\mathbb{E}(Z_kZ_l)\\
&= \sum_{k=1}^{n}\frac{\partial p_k}{\partial\theta_i}\frac{\partial p_l}{\partial\theta_j}\left(\mathbb{E}(Z_k^2)+\sum_{l\neq k}\mathbb{E}(Z_kZ_l)\right)\\
&= \sum_{k=1}^{n}\frac{\partial p_k}{\partial\theta_i}\frac{\partial p_l}{\partial\theta_j}\left(\text{var }(Z_k)+\sum_{l\neq k}\mathbb{E}(Z_k)\mathbb{E}(Z_l)\right)\\
&= \sum_{k=1}^{n}\frac{\partial p_k}{\partial\theta_i}\frac{\partial p_l}{\partial\theta_j}\frac{m_k}{p_k(1-p_k)}+0
\end{aligned}$$

as required. Notice that the proof does not rely on the form of the relationship between p_k and $\boldsymbol{\theta}$.

2. A orthopaediecs surgeon is investigating the functioning score of post-knee replacement surgery patients that underwent three different post-surgical intervention under three particpating physiotherapists. The data collected includes the response `FunctionScore` and two factor variables `Intervention` and `Therapist`. Unfortunately, a computer virus corrupted the output file. The corrupted output is as follows:

```
str(DataF)
'data.frame': 20 obs. of  3 variables:
$ FunctionScore: num  12.35 11.82 7.22 13.69 13.05 ...
$ Intervention : Factor w/ 3 levels "1","2","3": 1 1 3 2 2 1 1 1 2 2 ...
$ Therapist    : Factor w/ 3 levels "A","B","C": 3 1 1 3 3 2 2 2 1 1 ...
> summary(imodel)

Call:
lm(formula = FunctionScore ~ Intervention * Therapist, data = DataF)

Residuals:
Min      1Q  Median      3Q     Max
-0.8285 -0.1937  0.0000  0.3317  0.6139

Coefficients:
Estimate  Std. Error t value  Pr(>|t|)
(Intercept)               11.7250     0.3467  33.823 1.81e-12 ***
Intervention2             -3.2753     0.4903  -6.681 3.46e-05 ***
Intervention3             -4.1635     0.4475  -9.303 1.51e-06 ***
TherapistB                 2.5520     0.4246   6.011 8.79e-05 ***
TherapistC                 0.2161     0.4903   0.441    0.668
Intervention2:TherapistB  -0.8392     0.7354  -1.141    0.278
Intervention3:TherapistB  -0.2600     0.6169  -0.421    0.682
Intervention2:TherapistC   4.9021     0.6638   7.385 1.39e-05 ***
Intervention3:TherapistC   6.1056     0.7489   8.153 5.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: (d) on (a) degrees of freedom
Multiple R-squared:  0.9788,Adjusted R-squared:  0.9634
F-statistic: 63.49 on 8 and (a) DF,  p-value: 4.075e-08

> sum(imodel$residuals^2)
```

```
[1] 2.643836

summary(amodel)

Call:
lm(formula = FunctionScore ~ Intervention + Therapist, data = DataF)

Residuals:
Min       1Q   Median       3Q      Max
-2.91203 -0.68993  0.07554  1.02621  2.34695

Coefficients:
Estimate  Std. Error t value  Pr(>|t|)
(Intercept)    10.7563     0.7389  14.558 2.96e-10 ***
Intervention2  -1.7656     0.8253  -2.139 0.049253 *
Intervention3  -2.9097     0.8053  -3.613 0.002555 **
TherapistB      2.7526     0.8020   3.432 0.003705 **
TherapistC      3.6897     0.8361   4.413 0.000504 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.455 on (c) degrees of freedom
Multiple R-squared:  0.7455,Adjusted R-squared:  0.6777
F-statistic: 10.99 on 4 and (c) DF,  p-value: 0.0002298

> anova(amodel,imodel)
Analysis of Variance Table

Model 1: FunctionScore ~ Intervention + Therapist
Model 2: FunctionScore ~ Intervention * Therapist
Res.Df  RSS  Df   Sum of Sq      F    Pr(>F)
1     (c)  (e)
2     (a)  (b)  (f)    (g)         (h)   6.991e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Deduce the missing output (a) to (h).

Solution:

(a) Since there are 9 parameters in the reparameterised interaction model, (a) $= n-9 = 20-9 = 11$.

(b) Let e_i denote the i-th residual fort he reparameterised interaction model. Then (b) $= \sum e_i^2 = 2.643836$.

(c) Since there are 5 parameters in the reparameterised additive model, (c) $= n-5 = 20-5 = 15$.

(d) $s = \sqrt{\sum e_i^2/11} = 0.4902538$.

(e) For additive model, $SS_{res}^{H_0} = (s^{H_0})^2 \times (n-5) = 1.455^2 \times 15 = 31.75538$.

(f) DF of $SS_{res}^{H_0} - SS_{res}^{H_1} = 15 - 11 = 4$.

(g) $SS_{res}^{H_0} - SS_{res}^{H_1} = 31.75538 - 2.643836 = 29.11154$.

(h) F-statistic $= \{(SS_{res}^{H_0} - SS_{res}^{H_1})/4\}/\{0.4902538^2\} = 30.28051$.