MAST90104: A First Course in Statistical Learning

Week 10 Lab and Workshop

1. We revisit the pima dataset in Week 9. Remember that the data may be found in the the package faraway.

   (a) This question use a data set in package faraway. Load the package and read the help file (?pima) to get a description of the predictor and response variables, then use pairs and summary to perform simple graphical and numerical summaries of the data.
   Use the same set codes in Q2(a) Week 9 to remove observations with missing values.
   Solution:

```
> library(faraway)
> data(pima)
> View(pima)
> missing <- with(pima, missing <- glucose==0 | diastolic==0 | triceps==0 | bmi == 0)
> pima <- pima[!missing,]
```

   (b) Fit a probit regression model with test as the response and all the other variables as predictors. **Solution:**

```
> model <- glm(cbind(test, 1-test)~., family=binomial(link="probit"), data=pima)
> summary(model)

Call:
glm(formula = cbind(test, 1 - test) ~ ., family = binomial(link = "probit"),
data = pima)

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -5.6330061  0.5458534 -10.320   < 2e-16 ***
pregnant     0.0696354  0.0253570   2.746 0.006029 **
glucose      0.0219569  0.0026626   8.246   < 2e-16 ***
diastolic   -0.0055388  0.0060099  -0.922 0.356738
triceps      0.0042586  0.0085638   0.497 0.618992
insulin     -0.0007516  0.0005883  -1.277 0.201430
bmi          0.0500812  0.0135460   3.697 0.000218 ***
diabetes     0.6903347  0.2064695   3.344 0.000827 ***
age          0.0160165  0.0081559   1.964 0.049553 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 464.85  on 523  degrees of freedom
AIC: 482.85

Number of Fisher Scoring iterations: 5
```

   Answer the following questions using your fitted probit regression model.

   (c) Is the diastolic blood pressure significant in the regression model? Use your R output to evaluate its significance at 10% significance level.
   Solution: $\widehat{\beta}_{diastolic} = -0.0055388$. Wald's statistic $= \widehat{\beta}_{diastolic}/\text{SE}(\widehat{\beta}_{diastolic}) = -0.0055388/0.0060099 = -0.922$. Under H_0 of no effect, the test statistic follows $N(0,1)$. Since p-value $= 0.356738 > 0.10$, we don't reject H_0 and conclude that the diastolic blood pressure effect is not significant.

(d) Write down the formula for the fitted regression equation using your R output.
Solution:

$$\widehat{p} = \Phi(-5.6330+0.0696\texttt{pregnant}+0.0220\texttt{glucose}-0.0055\texttt{diastolic}+0.0043\texttt{triceps}-0.0008\texttt{insulin}+0.0501\texttt{bmi}+0.6903\texttt{diabetes}+0.0160\texttt{age})$$

(e) Predict the outcome for a woman with predictor values 1, 99, 64, 22, 76, 27, 0.25, 25 (same order as in the dataset). Give a 95% confidence interval for your prediction. Explain why the confidence is not symmetric about the estimated probability.

Solution:

```
> x <- predict(model, newdata = list(pregnant=1, glucose=99, diastolic = 64, triceps = 22,
insulin = 76, bmi=27, diabetes=.25, age=25), type="link", se.fit=TRUE)
> pnorm(c(x$fit-2*x$se.fit, x$fit, x$fit+2*x$se.fit))
1          1          1
0.01942196 0.03734525 0.06695185
```

Since the inverse link function Φ is non-linear, the confidence interval is not symmetric about the estimate probability.

2. In this question, we will generate simulated data using a probit model.

(a) Write a function in R with argument n that sets the random seed as `set.seed(n)` and generates independent draws $\{y_i\}_{i=1}^n$, where each y_i is drawn as

$$y_i \sim \text{Bin}(6, \Phi(-0.5 + 0.1x_{i1} - 0.2x_{i2}))$$

and each $\mathbf{x}_i = (x_{i1}, x_{i2})$ are drawn from a bivariate normal distribution with mean $\mathbf{0}$ and identity covariance matrix.
Solution:

```
SimulateData <- function(n){

set.seed(n)
X1 <- rnorm(n)
X2 <- rnorm(n)
beta.true <- c(-0.5,0.1,-0.2)
#pnorm equals to inverse link function for probit regression model
prob.true <- pnorm(c(cbind(1,X1,X2) %*% beta.true))
vy <- rbinom(n = n, size = 6, prob = prob.true)

return(data.frame(y=vy, x1=X1, x2=X2))

}
```

(b) Use the function in part (a) to generate a dataset of size $n = 30$.
Solution:

```
> genData <- SimulateData(30)
#head displays the first few rows of a data.frame object
> head(genData)
y         x1         x2
1 3 -1.2885182 -1.7252025
2 2 -0.3476894  0.6148607
3 2 -0.5216288  0.7268751
4 1  1.2734732 -0.0421902
5 3  1.8245206  0.2160018
6 1 -1.5113079  1.7697364
```

(c) Use the simulated dataset from part (b) to fit the binomial probit model:

$$y_i \sim \text{Bin}(6, \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))$$

Solution:

```
> ModelObj <- glm(cbind(y, 6-y) ~ ., data=genData, family = binomial(link="probit"))
> summary(ModelObj)

Call:
glm(formula = cbind(y, 6 - y) ~ ., family = binomial(link = "probit"),
data = genData)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.37424    0.10262  -3.647 0.000266 ***
x1           0.14217    0.10068   1.412 0.157914
x2          -0.17672    0.08288  -2.132 0.032987 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 21.300  on 29  degrees of freedom
Residual deviance: 14.956  on 27  degrees of freedom
AIC: 84.445

Number of Fisher Scoring iterations: 4
```

(d) Using your fitted model in part (c), construct a 90% confidence interval for

$$\Phi(\beta_0 - 0.5\beta_1 - 0.5\beta_2).$$

```
> xpred <- predict(ModelObj, newdata = list(x1=-0.5, x2=-0.5),
type = "link", se.fit=TRUE)
> pnorm(c(xpred$fit-qnorm(0.95)*xpred$se.fit, xpred$fit, xpred$fit
+qnorm(0.95)*xpred$se.fit))
1         1         1
0.3000107 0.3605586 0.4248260
```

1 Workshop questions

1. Suppose $Y_i, i = 1, \cdots, n$ are from a generalised linear model so they are independent from an exponential family:

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

with the parameter ϕ constant and supposed known but θ_i varies. Recall that

$$\begin{aligned} \mu &= \mathbb{E}Y = b'(\theta) \\ V(\mu) &= \operatorname{Var} Y = b''(\theta)a(\phi) \\ v &= b'' \circ (b')^{-1} \end{aligned}$$

and that there is a link function, g, so that $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ are the parameters of interest, $\mu_i = \mathbb{E}Y_i$ and $\mathbf{x}_i$ is a vector of explanatory variables (this is the ith row of the predictor matrix X). In answering the questions below, you will establish that the Newton-Raphson method with Fisher scoring is the same as the iteratively weighted least squares algorithm introduced in lectures.

(a) Write down the log likelihood as a function of $\boldsymbol{\beta}$ and show that its derivative, $U(\beta_j)$, with respect to β_j may be written as:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}.$$

Solution: The log likelihood is

$$\sum_{i=1}^{n} \log f(y_i;\theta_i,\phi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right]$$

so the derivative with respect to β_j is

$$\frac{\partial l(\boldsymbol{\beta};\mathbf{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{a(\phi)} \frac{\partial \theta_i}{\partial \beta_j}. \qquad (1)$$

Writing $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, $g(\mu_i) = \eta_i$ and $\mu_i = b'(\theta_i)$ so $\theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(\eta_i))$. Since $x = f^{-1}(y) \implies (f^{-1})'(y) = \frac{1}{f'(x)}$ applying the chain rule for differentiation twice gives

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{b''(\theta_i)} \frac{1}{g'(\mu_i)} x_{ij}.$$

Using the preamble formulae for mean and variance in equation 1 gives the required derivative of the log likelihood.

(b) Hence show that

$$Cov(U(\beta_j)U(\beta_k)) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{V(\mu_i)(g'(\mu_i))^2}.$$

Solution: The covariance is the expected value of the product $U(\beta_j)U(\beta_k)$ since both random variables have zero mean.

The terms in the multiplication of two sums are the sum of all the products. Both $U(\beta_j)$ and $U(\beta_k)$ are sums with n terms each. The terms where the indices in the sums are different all have expectation 0 since they are the expected value of a product of independent random variables each with zero mean. Hence

$$Cov(U(\beta_j)U(\beta_k)) = \sum_{i=1}^{n} E\left(\frac{(y_i - \mu_i)^2}{V^2(\mu_i)} \frac{x_{ij}x_{ik}}{(g'(\mu_i))^2}\right)$$

giving the required equation since $\operatorname{Var}(y_i) = E((y_i - \mu_i)^2)$.

(c) Find the Fisher information and show that it is $X^T W(\boldsymbol{\beta})X$ where $W(\boldsymbol{\beta})$ is a diagonal matrix whose ith diagonal entry is

$$\frac{1}{V(\mu_i)(g'(\mu_i))^2}.$$

Solution: The Fisher information matrix is defined to the matrix whose entries are

$$Cov(U(\beta_j)U(\beta_k)), j, k = 1, \cdots, n$$

.

If D is a diagonal matrix with diagonal entries $d_i, i = 1, \cdots, n$, then the jth row of X^T is the row vector with entries $x_{ij}d_i, i = 1, \cdots, n$. Taking the dot product of this vector with the kth column of X gives $\sum_{i=1}^n x_{ij}x_{ik}d_i$. This expression is the (j, k) entry of $X^T DX$.

Taking D to be the diagonal matrix $W(\boldsymbol{\beta})$ and using part (b) gives the required expression for the Fisher information matrix.

2. Suppose that students answer questions on a test and that a specific student has an aptitude T. A particular question might have difficulty d_i and the student will get the answer correct only if $T > d_i$. Consider d_i fixed and $T \sim N(\mu, \sigma^2)$, then the probability that a randomly selected student will get the answer wrong is $p_i = \mathbb{P}(T < d_i)$.

Show how you might model this situation using a probit regression model.

Solution: We have

$$\begin{aligned} p_i &= \mathbb{P}(T < d_i) \\ &= \mathbb{P}\left(\frac{T-\mu}{\sigma} < \frac{d_i - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{1}{\sigma}d_i - \frac{\mu}{\sigma}\right) \end{aligned}$$

which is in the form of a probit regression model with predictor variable d, $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$.

3. Show that the Gamma density, f, in the form

$$f(y; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha y^{\alpha-1}e^{-\lambda y}$$

is an exponential family with $\theta = -\frac{\lambda}{\alpha}, \phi = \frac{1}{\alpha}$. Identify the functions a, b, c and find the mean and variance functions as functions of θ.

Solution: Notice that $\theta/\phi = \theta\alpha = -\lambda$, so

$$\log f(y; \theta, \phi) = \frac{y\theta + \log(-\theta)}{\phi} + \frac{1-\phi}{\phi}\log(y) - \frac{\log(\phi)}{\phi} - \log\left(\Gamma(1/\phi)\right)$$

So the functions are $a(\phi) = \phi, b(\theta) = -\log(-\theta), c(y, \phi) = \log(y)(1-\phi)/\phi - \log(\phi)/\phi - \log\left(\Gamma(1/\phi)\right)$.

This gives $b'(\theta) = -1/\theta, b''(\tau) = \theta^{-2}$ giving $E(Y) = -1/\theta (= \alpha/\lambda)$ and $var(Y) = b''(\theta)a(\phi) = \theta^2\phi = \theta^2/\alpha = (\frac{\alpha}{\lambda^2})$.

Note that the formulas for the mean and variance are, of course, the same as those derived in MAST90105 using moment generating functions (or the change parameters trick). Also note that b' is self inverse so the canonical link function is the negative inverse.

4. Show that the inverse Gaussian density, f, in the form

$$f(y; \mu, \lambda) = \frac{\lambda}{\sqrt{2\pi y^3}}e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$$

is an exponential family with $\theta = \frac{-1}{2\mu^2}, \phi = \frac{1}{\lambda}$. Identify the functions a, b, c and find the mean and variance functions as functions of μ, λ.

Solution: Writing the density function in terms of the defined parameters, θ, ϕ

$$\log f(y; \theta, \phi) = \frac{y\theta - (-\sqrt{-2\theta})}{\phi} + (-\log(\phi) - \log(2\pi)/2 - \frac{3}{2}\log(y) - \frac{1}{2\phi y}$$

.

So the functions are $a(\phi) = \phi, b(\theta) = -\sqrt{-2\theta}, c(y, \phi) = -\log(\phi) - \log(2\pi)/2 - \frac{3}{2}\log(y) - \frac{1}{2\phi y}$.

This gives $b'(\theta) = 1/\sqrt{-2\theta}, b''(\tau) = (-2\theta)^{-3/2}$ giving $E(Y) = 1/\sqrt{-2\theta} = \mu$ and $var(Y) = b''(\tau)a(\phi) = \theta^{-2}\phi = \mu^3/\lambda$.

Note that the canonical link is half the negative inverse of the square.