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
```