Math-661: Assignment 3 – Sample Solution

1. **Exercise 1 (Agresti 5.20)**

   Let $y_i, i = 1, \dots, N$, denote N independent binary random variables.

$$y_i \sim \text{ Binomial}(1, \pi_i) \qquad i = 1, \dots, N$$

The likelihood function is given by

$$\mathcal{L}(\boldsymbol{\pi}) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

and the log-likelihood function is

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{N} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)$$

(a) **Derive the log-likelihood for the probit model $\Phi^{-1}[\pi(\boldsymbol{x}_i)] = \sum_j \beta_j x_{ij}$.**

For the probit model

$$\Phi^{-1}(\pi_i) = \sum_j \beta_j x_{ij} \qquad \Rightarrow \qquad \pi_i = \Phi\left(\sum_j \beta_j x_{ij}\right)$$

so the log-likelihood becomes

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \log\left(\frac{\Phi(\sum_j \beta_j x_{ij})}{1-\Phi(\sum_j \beta_j x_{ij})}\right) + \log\left(1 - \Phi\left(\sum_j \beta_j x_{ij}\right)\right)$$

(b) **Show that the log-likelihood equations for the logistic and probit regression models are**

$$\sum_{i=1}^{N}(y_i - \hat{\pi}_i) z_i x_{ij} = 0, \qquad j = 1, \dots, p,$$

where $z_i = 1$ for the logistic case and $z_i = \phi(\sum_j \hat{\beta}_j x_{ij})/[\hat{\pi}_i(1-\hat{\pi}_i)]$ for the probit case.

The score/log-likelihood equation for the probit regression model is

$$\begin{aligned}\frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) \quad = \quad & \sum_{i=1}^{N}\left[y_i \left(\frac{1-\Phi(\sum_j \beta_j x_{ij})}{\Phi(\sum_j \beta_j x_{ij})}\right) \times \frac{\phi(\sum_j \beta_j x_{ij}) x_{ij}\left[1-\Phi(\sum_j \beta_j x_{ij})\right] + \phi(\sum_j \beta_j x_{ij}) x_{ij} \Phi(\sum_j \beta_j x_{ij})}{\left[1-\Phi(\sum_j \beta_j x_{ij})\right]^2} \right. \\ & \left. + \left(\frac{1}{1-\Phi(\sum_j \beta_j x_{ij})}\right) \times \left(-\phi\left(\sum_j \beta_j x_{ij}\right) x_{ij}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \left[y_i \frac{x_{ij}\phi(\sum_j \beta_j x_{ij})}{\Phi(\sum_j \beta_j x_{ij}) \left[1 - \Phi(\sum_j \beta_j x_{ij})\right]} - \frac{x_{ij}\phi(\sum_j \beta_j x_{ij})}{1 - \Phi(\sum_j \beta_j x_{ij})} \right] \\
&= \sum_{i=1}^{N} \left[y_i \frac{x_{ij}\phi(\sum_j \beta_j x_{ij})}{\pi_i(1-\pi_i)} - \frac{x_{ij}\phi(\sum_j \beta_j x_{ij})}{1-\pi_i} \right] \\
&= \sum_{i=1}^{N} (y_i - \pi_i) \frac{x_{ij}\phi(\sum_j \beta_j x_{ij})}{\pi_i(1-\pi_i)} = 0
\end{aligned}
$$

Thus, the score equation for the probit regression model is

$$
\sum_{i=1}^{N} (y_i - \hat{\pi}_i) z_i x_{ij} = 0 \qquad \text{where} \qquad z_i = \frac{\phi(\sum_j \beta_j x_{ij})}{\pi_i(1-\pi_i)}
$$

The logit link is given by

$$
\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_j \beta_j x_{ij} \qquad\qquad \pi_i = \frac{\exp(\sum_j \beta_j x_{ij})}{1 + \exp(\sum_j \beta_j x_{ij})}
$$

and the log-likelihood for the logistic regression model is

$$
l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[y_i \sum_j \beta_j x_{ij} - \log\left(1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right) \right]
$$

so the score equation for the logistic regression model is

$$
\frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[y_i x_{ij} - \frac{x_{ij} \exp(\sum_j \beta_j x_{ij})}{1 + \exp(\sum_j \beta_j x_{ij}} \right] = \sum_{i=1}^{N} (y_i x_{ij} - \pi_i x_{ij}) = 0
$$

Thus, the score equation for the logistic regression model is

$$
\sum_{i=1}^{N} (y_i - \hat{\pi}_i) x_{ij} = 0
$$

2. **Exercise 2 (based on Agresti 5.32)**

   For the horseshoe crab dataset (`Crabs.txt`**), let $y = 1$ if a female crab has at least one satellite, and let $y = 0$ if a female crab does not have any satellite.**

```
# create indicator variable y
> Crabs$y = as.numeric(Crabs$satellite>0)
> attach(Crabs)
```

   (a) **Fit a main-effects logistic model using color and weight as explanatory variables.**

```
# color is a categorical variable with 4 levels
> table(color)
color
 1  2  3  4
12 95 44 22

> fit.main = glm(y ˜ weight+as.factor(color), family=binomial)
> summary(fit.main)

Call:
glm(formula = y ˜ weight + as.factor(color), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1908  -1.0144   0.5101   0.8683   2.0751

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.2572     1.1985  -2.718  0.00657 **
weight              1.6928     0.3888   4.354 1.34e-05 ***
as.factor(color)2   0.1448     0.7365   0.197  0.84410
as.factor(color)3  -0.1861     0.7750  -0.240  0.81019
as.factor(color)4  -1.2694     0.8488  -1.495  0.13479
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 188.54  on 168  degrees of freedom
AIC: 198.54

Number of Fisher Scoring iterations: 4
```

   i. **Interpret the regression coefficients.**

   Controlling for color, a 1-unit increase in weight is associated with a 1.7 increase in log-odds of having at least one satellite.

Since we fail to reject $H_0 : \beta_j = 0$ for the effects of color, the regression coefficient estimates we get are just random deviations from 0 and it does not make sense to interpret them. If we still go ahead and interpret the coefficients of color (as an exercise), we would say, for example, that adjusting for weight, the log-odds of having at least one satellite for a female crab with color=2 is higher by 0.145 compared to a female crab with color=1.

ii. **Show how to conduct inference about the color and weight effects (i.e., evaluate statistical significance).**

To test $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ using Wald test,

$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

The test statistic to assess the effect of weight is

$$z = \frac{0.1448}{0.7365} = 4.354$$

and the corresponding p-value

$$2 \times P(Z \geq 4.354) = 1.337 \times 10^{-5}$$

```
> 2*(1-pnorm(4.354))
[1] 1.336757e-05
```

Since the p-value is very small, we reject H_0 and conclude that there is a significant effect of weight on the probability of having at least one satellite, after controlling for color.

To evaluate the effect of color = 2 compared to color = 1, the test statistic is

$$z = \frac{1.6928}{0.3888} = 0.1966056$$

and the corresponding p-value is

$$2 \times P(Z \geq 0.197) = 0.844$$

```
> 2*(1-pnorm(0.1966056))
[1] 0.8441362
```

Since the p-value is large, we fail to reject H_0 and conclude that there is no evidence to suggest that the odds of having at least one satellite are different for color=2 compared to color=1, after controlling for weight.

If we want to evaluate the effect of color (across all levels),

$$H_0 : \beta_{c2} = \beta_{c3} = \beta_{c4} = 0$$

we can use a likelihood ratio test comparing the model above to one with only weight. This can be achieved by evaluating the change in deviance, which will follow an approximate chi-square distribution with $df = 5 - 2 = 3$:

```
> weight.fit = glm(y ~ weight, family=binomial)
> deviance(weight.fit) - deviance(fit.main)
[1] 7.194895
> 1-pchisq(deviance(weight.fit) - deviance(fit.main), 3)
[1] 0.06593852
```

At $\alpha = 0.05$, we fail to reject H_0 and conclude that there is no evidence (or marginal evidence) that color is associated with the probability of having at least one satellite, adjusting for weight.

(b) **Allow interaction between color and weight in their effects on y.**

```
> fit.inter = glm(y ~ weight*as.factor(color), family=binomial)
> summary(fit.inter)

Call:
glm(formula = y ~ weight * as.factor(color), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0875  -0.8766   0.5412   0.8399   1.9421

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.6203     4.8909  -0.331    0.740
weight                     1.0483     1.8929   0.554    0.580
as.factor(color)2         -0.8320     5.0311  -0.165    0.869
as.factor(color)3         -6.2964     5.5165  -1.141    0.254
as.factor(color)4          0.4335     5.4046   0.080    0.936
weight:as.factor(color)2   0.3613     1.9559   0.185    0.853
weight:as.factor(color)3   2.7065     2.2284   1.215    0.225
weight:as.factor(color)4  -0.8536     2.1551  -0.396    0.692

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 181.66  on 165  degrees of freedom
AIC: 197.66

Number of Fisher Scoring iterations: 5
```

i. **Interpret the regression coefficients.**

None of the regression coefficients achieve statistical significance, so it does not make sense to interpret them since the estimates we get are just random variations from 0.
However, if we still go ahead and interpret the regression coefficients (as an exercise), we would say:

- For a female crab with color=1, each additional 1-unit increase in weight is associated with a 1.05 increase in the log-odds of having at least one satellite.

- For a female crab with color=2, each additional 1-unit increase in weight is associated with a 1.41 (1.0483+0.3613) increase in the log-odds of having at least one satellite.
- For a female crab with color=3, each additional 1-unit increase in weight is associated with a 3.75 (1.0483+2.7065) increase in the log-odds of having at least one satellite.
- For a female crab with color=4, each additional 1-unit increase in weight is associated with a 0.19 (1.0483-0.8536) increase in the log-odds of having at least one satellite.

ii. **Test whether this model provides a significantly better fit compared to the main-effects model.**

$$H_0 : \text{model with main effects fits as well as model with interaction effects}$$

We can use a likelihood ratio test to compare this model to the main effects model, which will have an approximate chi-square distribution with $df = 8 - 5 = 3$. This is equivalent to evaluating the change in deviance between the two models:

```
> deviance(fit.main)-deviance(fit.inter)
[1] 6.886003
> 1-pchisq(deviance(fit.main)-deviance(fit.inter), 3)
[1] 0.07562139
```

We fail to reject H_0 at $\alpha = 0.05$, thus the model with interaction effects does not provide a better fit compared to the model with main effects.

3. **Exercise 3: Survival of the Donner Party**

In 1846, a group of 87 people (called the Donner Party) were headed west from Springfield, Illinois, to California. The leaders attempted a new route through the Sierra Nevada and were stranded throughout the winter. The harsh weather conditions and lack of food resulted in the death of many people within the group. Social scientists have used the data to study the theory that females are better able than males to survive harsh conditions. The data are saved under `Donner.txt`.

(a) **Create a logistic regression model using gender and age as predictors and provide the equation of the estimated model.**

```
> donner.fit = glm(Survived ~ Male.Gender + Age, family=binomial)
> summary(donner.fit)

Call:
glm(formula = Survived ~ Male.Gender + Age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9152  -1.0340   0.6291   1.0385   1.6698

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  1.69486    0.52039   3.257  0.00113 **
Male.Gender -1.19255    0.49317  -2.418  0.01560 *
Age         -0.03503    0.01611  -2.175  0.02964 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.06  on 85  degrees of freedom
Residual deviance: 105.50  on 83  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 111.5

Number of Fisher Scoring iterations: 4
```

The logistic regression model is given by

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.695 - 1.193\ Male - 0.035 Age$$

(b) **Interpret the regression coefficients.**

All the regression coefficients are statistically significant, so it makes sense to interpret them.

- Controlling for age, the log-odds of survival are 1.193 lower for males compared to females. In other words, controlling for age, the odds of survival for males is $\exp(-1.19255) = 0.303$ times lower compared to females.
- Holding gender fixed, a one-year increase in age is associated with a 0.035 lower log-odds of survival. In other words, given the same gender, for every additional one-year increase in age, the odds of survival is $\exp(-0.3503) = 0.966$ times lower.

(c) **Estimate the survival probability of a 20-year old female (show your calculation).**

We have

$$\hat{\pi}_i = \frac{\exp(\sum_j \hat{\beta}_j x_{ij})}{1+\exp(\sum_j \hat{\beta}_j x_{ij})} = \frac{\exp(1.695 - 1.193(0) - 0.035(20))}{1+\exp(1.695 - 1.193(0) - 0.035(20))} = 0.73$$

```
> exp(1.69486-0.03503*20)/(1+exp(1.69486-0.03503*20))
[1] 0.7299285
```

Using R:

```
> predict(donner.fit, newdata=data.frame(Male.Gender=0, Age=20), type="response")
        1
0.7299373
```

Thus, a 20-year old female has a predicted survival probability of 0.73.

(d) **Explain why the deviance or Pearson goodness-of-fit tests are not appropriate.**

There is only one individual falling in many of the covariate patterns made up of age and gender, because of the continuous variable age. Therefore the sample size requirement needed to use the chi-square approximation for the deviance and the Pearson goodness-of-fit tests is not satisfied, making the tests inappropriate.

(e) **Assess the model goodness-of-fit.**

We can the Hosmer-Lemeshow goodness-of-fit test to assess the model fit.

$$H_0 : \text{the model fits the data well.}$$

Using the function `hoslem.test()` from the R package `ResourceSelection`, we get

```
 > hoslem.test(donner.fit$y, fitted(donner.fit))

Hosmer and Lemeshow goodness of fit (GOF) test

data:  donner.fit$y, fitted(donner.fit)
X-squared = 12.956, df = 8, p-value = 0.1134
```

We fail to reject H_0 at $\alpha = 0.05$. Thus, there is not sufficient evidence to suggest that the model does not provide adequate fit to the data. However given the marginally significant p-value, we should try to find a better fitting model.