Michael Leibert
Math 661
Homework 3

1. **Exercise 1 (Agresti 5.20)**
   Let $y_i, i = 1, \ldots, N$, denote N independent binary random variables.

   (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, we use the standard cdf of the normal distribution.

$$P(y_i = 1) = \pi_i = \Phi\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$$

   The likelihood for a bernoulli random variable:

$$\begin{aligned}
f(y_i) &= \prod_{i=1}^{n} \left[P\left(y_i = 1\right)\right]^{y_i} \Big[P\left(y_i = 0\right)\Big]^{1-y_i} \\
&= \prod_{i=1}^{n} \Big[\Phi\left(\eta_i\right)\Big]^{y_i} \Big[1 - \Phi\left(\eta_i\right)\Big]^{1-y_i} \\
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \log\Big[\Phi\left(\eta_i\right)\Big] + \left(1 - y_i\right) \log\Big[1 - \Phi\left(\eta_i\right)\Big] \\
&= \sum_{i=1}^{n} y_i \log\Big[\Phi\left(\eta_i\right)\Big] + \log\Big[1 - \Phi\left(\eta_i\right)\Big] - y_i \log\Big[1 - \Phi\left(\eta_i\right)\Big] \\
&= \sum_{i=1}^{n} y_i \log\left(\frac{\Phi\left(\eta_i\right)}{1 - \Phi\left(\eta_i\right)}\right) + \log\Big[1 - \Phi\left(\eta_i\right)\Big]
\end{aligned}$$

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

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

   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.

   Likelihood score equations:

$$\sum_{i=1}^{n} \frac{\left(y_i - \pi_i\right)\ x_{ij}}{Var\left(Y_i\right)} \frac{\partial \pi_i}{\partial \eta_i}.$$

   For Logistic $\eta_i = \theta_i$ and $\pi_i = \dfrac{\exp\left(\eta_i\right)}{1 + \exp\left(\eta_i\right)}$.

$$\begin{aligned}
\frac{\partial}{\partial \eta_i} \frac{\exp(\eta_i)}{1+\exp(\eta_i)} &= \frac{\exp(\eta_i)}{\left[1+\exp(\eta_i)\right]^2} \\
&= \pi_i \frac{1}{1+\exp(\eta_i)} \\
&= \pi_i \frac{1+\exp(\eta_i)-\exp(\eta_i)}{1+\exp(\eta_i)} \\
&= \pi_i \left(1 - \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \\
&= \pi_i (1-\pi_i)
\end{aligned}$$

Likelihood score equation for logistic:

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)\ x_{ij}}{\cancel{Var(Y_i)}} \cancel{\hat{\pi}_i(1-\hat{\pi}_i)} = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)\, x_{ij}\ z_i \quad \text{where } z_i = 1$$

For Probit, $\pi_i = \Phi\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) = \Phi(\eta_i)$.

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} \Phi(\eta_i) = \phi(\eta_i)$$

Likelihood score equation for logistic:

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)\ x_{ij}}{Var(Y_i)} \phi(\eta_i) = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)\, x_{ij}\ z_i \quad \text{where } z_i = \frac{\phi(\eta_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} = \frac{\phi\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

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.

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

```
tail(crabs)
##     weight color y
## 168  2.175     3 1
## 169  2.750     3 1
## 170  3.275     3 1
## 171  2.625     1 0
## 172  2.625     4 0
## 173  2.000     2 0
str(crabs)
## 'data.frame': 173 obs. of  3 variables:
##  $ weight: num  3.05 1.55 2.3 2.1 2.6 2.1 2.35 1.9 1.95 2.15 ...
##  $ color : Factor w/ 4 levels "1","2","3","4": 2 3 1 3 3 2 1 3 2 3 ...
##  $ y     : num  1 0 1 0 1 0 0 0 0 0 ...
```

```
crab.glm<-glm( y~.,data=crabs ,family="binomial" )
summary(crab.glm)
```

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = crabs)
##
## 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 ***
## color2        0.1448     0.7365   0.197  0.84410
## color3       -0.1861     0.7750  -0.240  0.81019
## color4       -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, for every one unit increase in weight, the log odds of having at least one satellite increases by 1.6928.

The regression coefficients for color are not significant so we cannot interpret them.

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

We can test $H_0 : \beta_1 = 0$ using the Wald test:

$$z = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} = \frac{1.6928}{0.3888} = 4.353582$$

and yields a p-value $= 2 \cdot P(Z \geq 4.353582) = 0.0000133930966$. Thus, we reject H_0. There is strong evidence of an association between weight of the crab and the absence/presence of a satellite.

We can also derive a confidence interval for β_1. A 95% Wald confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm 1.959 \cdot SE\left(\hat{\beta}_1\right) = 1.6928 \pm 1.959 \cdot 0.3888 = (0.93, 2.45)$$

We are 95% confident that the log-odds of the presence of a satellite is expected to be at least 0.93 and at most 2.45 higher for heavier crabs, on average, controlling for the crab's color. We note that the confidence interval does not contain zero, which is another way to determine we will reject H_0.

```
library(MASS)
coef(crab.glm)[2] - qnorm(.975)*sqrt(diag(vcov(crab.glm)))[2]; coef(crab.glm)[2] +
qnorm(.975)*sqrt(diag(vcov(crab.glm)))[2]
```

```
##   weight
## 0.930724
##   weight
## 2.454931
```

```
confint.default(crab.glm)
```

```
##                 2.5 %     97.5 %
## (Intercept) -5.606109 -0.9082301
## weight       0.930724  2.4549314
## color2      -1.298678  1.5883425
## color3      -1.705046  1.3327764
## color4      -2.933126  0.3942691
```

```
1-pchisq(7.194895 , 3)
```

```
## [1] 0.06593853
```

To test the significance of color, controlling for weight we will test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. The likelihood-ratio statistic is $G^2 = -2(\mathcal{L}_0 - \mathcal{L}_1)$ with 3 degrees of freedom.

$$\begin{aligned} G^2 &= -2(\mathcal{L}_0 - \mathcal{L}_1) \\ &= -2\big(- 97.86857488 - (-94.27112714)\big) \\ &= 7.194895483 \end{aligned}$$

This is the same as difference between the residual deviances: $195.73715 - 188.54225 = 7.1949$.

The test statistic yields a p-value $= P\left(\chi^2_3 \geq 7.1949\right) = 0.06593853$. The p-value is right on the border for the 0.05 cutoff. Considering it is above the $\alpha = 0.05$ and, the confidence intervals listed above all contain 0 for the colors, we will accept $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. There is weak to no evidence of an association between color of the crab and the absensce/presnce of a satellite.

```
crabby.glm<-glm(y~weight,data=crabs ,family="binomial" )
summary(crabby.glm)
```

```
##
## Call:
## glm(formula = y ~ weight, family = "binomial", data = crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1108  -1.0749   0.5426   0.9122   1.6285
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
## weight        1.8151     0.3767   4.819 1.45e-06 ***
## ---
## 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: 195.74  on 171  degrees of freedom
```

```
## AIC: 199.74
##
## Number of Fisher Scoring iterations: 4
-2 * ( logLik(crabby.glm) - logLik(crab.glm) )
## 'log Lik.' 7.194895 (df=2)
195.73715 - 188.54225
## [1] 7.1949
1-pchisq(7.194895 , 3)
## [1] 0.06593853
```

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

```
crabI.glm<-glm( y~weight*color,data=crabs ,family="binomial" )
summary(crabI.glm)

##
## Call:
## glm(formula = y ~ weight * color, family = "binomial", data = crabs)
##
## 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
## color2         -0.8320     5.0311  -0.165    0.869
## color3         -6.2964     5.5165  -1.141    0.254
## color4          0.4335     5.4046   0.080    0.936
## weight:color2   0.3613     1.9559   0.185    0.853
## weight:color3   2.7065     2.2284   1.215    0.225
## weight:color4  -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 are significant so we cannot interpret them.

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

We can do several things to test whether the interaction models provides a better fit compared to the main-effects one. First we can do a likelihood ratio test (LRT). We note that the the main-effects model is nested within the interaction one. So we test H_0: The smaller model fits well.

The test statistic is $\Delta G^2 = G_0^2 - G_1^2 \rightarrow \chi_v^2$. Where G_0^2 is the deviance from the main-effects model and G_1^2 is the deviance from the interaction model. Where $v = 3$, is the difference in the number of parameters between the two models.

```
1-pchisq(188.54-181.66 , 168-165)
## [1] 0.07582254
```

This is a pretty low p-value for H_0: smaller model fits well. The .05 threshold is too low, and the .07 does not show that the nested model does better than the interaction model.

We can also look at AIC and BIC criteria for model comparison.

```
AIC(crab.glm)
## [1] 198.5423
AIC(crabI.glm)
## [1] 197.6563
```

AIC is very close, but it is suggesting the larger model as well.

```
BIC(crab.glm)
## [1] 214.3087
BIC(crabI.glm)
## [1] 222.8826
```

However, for BIC the penalty for the larger model is suggesting the smaller one.

I would say overall it is pretty close, but favoring the LRT I would conclude going with the interaction model. Although I do not think it is accurate to say that the interaction model is significantly better than the main effects model.

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.

```
tail(donner)
##    Male.Gender Age Survived
## 82           1  35        0
## 83           1  23        1
## 84           1  24        0
## 85           0  25        1
## 86           1  NA        0
## 87           0  20        1
```

```
donner.glm<-glm(Survived~.,data=donner,family=binomial )
summary(donner.glm)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = donner)
##
## 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.Gender1 -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
```

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.69486 - 1.19255x_{i1} - 0.03503x_{i2}$$

(b) Interpret the regression coefficients.

Controlling for gender, for every one year increase in age, the log odds of surviving decreases by 0.03502802 (increases by -0.03502802).

Controlling for age, males have a 1.19255 lower log odds than women for surviving.

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

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 1.69486 - 1.19255(1) - 0.03503(20)$$

$$\hat{\pi} = \frac{\exp(0.9943046)}{1+\exp(0.9943046)}$$

$$\hat{\pi} = 0.7299373$$

```
eta<- sum( coef(donner.glm) * c(1,0,20) )
exp(eta)/(1+exp(eta))
```

```
## [1] 0.7299373
```

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

These data are not grouped so the deviance or Pearson goodness-of-fit tests are not appropriate.

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

We conduct the Hosmer-Lemeshow goodness of fit test to assess the model goodness-of-fit. We test H_0: the current model fits well.

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 3.3.3
```

```
## ResourceSelection 0.3-2   2017-02-28
```

```
res<-hoslem.test(donner.glm$y,fitted(donner.glm))
res
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  donner.glm$y, fitted(donner.glm)
## X-squared = 12.956, df = 8, p-value = 0.1134
```

```
cbind(res$observed,res$expected)
```

```
##               y0 y1    yhat0    yhat1
## [0.145,0.358]  6  3 6.841815 2.158185
## (0.358,0.408] 10  4 8.558368 5.441632
## (0.408,0.425]  3  1 2.309556 1.690444
## (0.425,0.512]  3  5 4.071472 3.928528
## (0.512,0.547]  4  5 4.210162 4.789838
## (0.547,0.59]   4  5 3.787288 5.212712
## (0.59,0.694]   4  4 2.923714 5.076286
## (0.694,0.769]  0  8 2.158702 5.841298
## (0.769,0.82]   0  8 1.657358 6.342642
## (0.82,0.84]    4  5 1.481564 7.518436
```

This p-value of 0.1134 is non-significant at the 0.05 level so there should be no evidence that the model is fitting poorly. However, because the p-value is near the threshold there may be evidence that the model is not a great fit. With caution, we declare the model does appear adequate for these data.