STAA552 – Solution 5

1. Scientists agree that preserving certain habitats in their natural state is necessary to slow the accelerating rate of species extinctions. But, they are undecided on how to construct such reserves. Specifically, is it better to have many small reserves or a few large ones? Observe the following study.

Island	Area (km^2)	Species at risk	Extinctions
Ulkokrunni	185.80	75	5
Maakrunni	105.80	67	3
Ristikari	30.70	66	10
Isonkivenletto	8.50	51	6
Heitakraasukka	4.80	28	3
Kraasukka	4.50	20	4
Lansiletto	4.30	43	8
Pihlajakari	3.60	31	3
Tyni	2.60	28	5
Tasasenletto	1.70	32	6
Raiska	1.20	30	8
Pohjanletto	0.70	20	2
Toro	0.70	31	9
Luusiletto	0.60	16	5
Vatunginletto	0.40	15	7
Vatunginnokka	0.30	33	8
Tiirakari	0.20	40	13
Ristikarenletto	0.07	6	3

Table 1: Island area, number of bird species present in 1949, and number of those not present in 1959; Krunnit Islands study.

Our goal is to fit a binomial regression with logit link for the probability of extinction as a function of Area. Use any software you wish.

a. Plot the logit(proportion of extinction) as a function of Area. Why does a linear regression model not seem appropriate?

b. Plot the logit(proportion of extinction) as a function of log(Area). Does a linear regression model seem appropriate?

c. After making a log transformation to Area, fit a logistic regression curve to the data. What are $\hat{\beta}_0$ (intercept), $\hat{\beta}_1$ (slope), the estimated standard errors of both, and the confidence intervals for both?

d. Perform the goodness of fit 'Drop-in-Deviance' test for this model fit versus the intercept only model. What are the null and alternate hypotheses this statistic is testing? What is the general form for this test statistic in this case (that is, it should involve writing terms like L_M from lecture)? What is the associate p-value? Is this a valid test to use in the situation and why? Compare this p-value and Chi-squared statistic to the ones found in c.

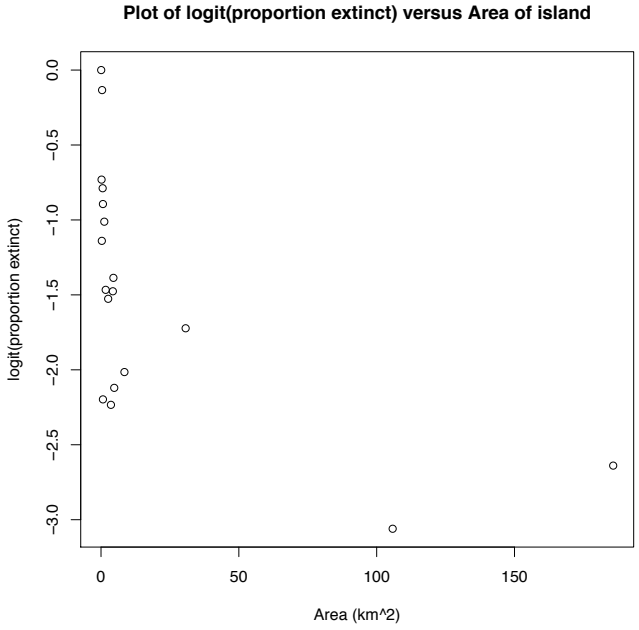

Figure 1: Problem 4a. The data linear trend. There does seem to be substantial need for a transformation of the explanatory variable.

e. Interpret the estimated coefficient on log(Area). Get the fitted probability for Pihlajakari and for Maakrunni. Which island would we predict to have a greater rate of extinction?

f. It seems possible that there is a bit of a quadratic slope in the data as well. Get the model fit with the additional quadratic term. Give a formal test for the inclusion of a quadratic term using the drop in deviance.

g. What conclusions might you draw from this study? What is the scope of this inference (that is, can your conclusions be extrapolated to other situations with different islands or habitats)?

h. Look at the overall Deviance/Pearson Chi Squared goodness of fit test. What are the null and alternate hypotheses this statistic is testing? What is the general form for this test statistic in this case (that is, it should involve writing terms like L_M from lecture)? What does this test indicate as far as model fit? If warranted, go through the steps related to overdispersion[1].

---

a. See Figure 1.

b. See Figure 2.

c.
```
area    = c(185.8, 105.8,30.7,8.5,4.8,4.5,4.3,3.6,2.6,1.7,1.2,.7,.7,.6,.4,.3,.2,.07)
species = c(75,67,66,51,28,20,43,31,28,32,30,20,31,16,15,33,40,6)
extinct = c(5,3,10,6,3,4,8,3,5,6,8,2,9,5,7,8,13,3)
```

[1]Important: The steps in the question are out of order for doing an actual analysis. My motivation is I didn't want some people to do this problem with overdispersion, others without it. Remember, always look for overdispersion first.

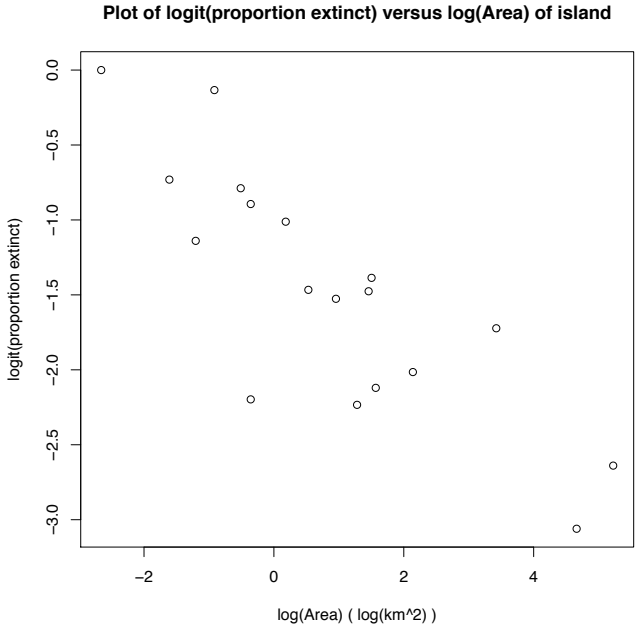

Figure 2: Problem 4b. After the log transformation of area, the trend appears to be much more linear.

```
prop    = extinct/species
designMatrix = cbind(extinct,species-extinct)
island = cbind(area,species,extinct)
colnames(island) = c("area","species","extinctions")

out = glm(designMatrix~log(area),family=binomial(logit))
Call:
glm(formula = designMatrix ~ log(area), family = binomial(logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.71726  -0.67722   0.09726   0.48365   1.49545

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
log(area)   -0.29710    0.05485  -5.416 6.08e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.338  on 17  degrees of freedom
Residual deviance: 12.062  on 16  degrees of freedom
```

```
AIC: 75.394

Number of Fisher Scoring iterations: 4
```

d.
```
> #####
> ## 4d
> #####
> print(pchisq(out$null.deviance - out$deviance,df=1,lower=F))
[1] 7.994246e-09
##P-value for intercept + slope model vs.
##intercept only model
```

e. The estimated coefficient on log(area) is -0.297. This means that a unit increase in the explanatory variable leads to a multiplicative increase in the odds of $\exp\{-0.297\} = 0.743$. If the explanatory variable is $\log(X)$, then the interpretation is more conveniently presented as a multiplicative change in X. For instance, for each doubling of X, $\log(X)$ increases by $\log(2)$. This means that the log odds change by $\beta_1 \log(2)$ units. Hence, the odds change by 2^{β}.

So, each doubling of island size is associated with a change in the odds of extinction by a multiplicative factor of $2^{-2.97} = 0.813$. For example, the estimated odds of extinction for a species on an island of area 2A is 81.3% of the odds of extinction of an island of area A.

f. Here is the output:

```
> outQuad = glm(designMatrix~log(area)+ I(log(area)^2),family=binomial(logit))
> summary(outQuad)

Call:
glm(formula = designMatrix ~ log(area) + I(log(area)^2), family = binomial(logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.70543  -0.64658   0.02695   0.49182   1.47740

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.21118    0.12971  -9.337  < 2e-16 ***
log(area)      -0.31780    0.09047  -3.513 0.000444 ***
I(log(area)^2)  0.00699    0.02430   0.288 0.773642
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.338  on 17  degrees of freedom
Residual deviance: 11.979  on 15  degrees of freedom
```

Here is conclusion:

```
> #####
> ## 4e
> #####
```

```
> outQuad = glm(designMatrix~log(area)+ I(log(area)^2),family=binomial(logit))
> print(outQuad$deviance - out$deviance)
[1] -0.08231
> print(pchisq(out$deviance - outQuad$deviance,df=1,lower=F))
[1] 0.774191
 ##P-value for inclusion of quadratic term
```

g. Since the data are observational, confounding variables associated with both island size and the event of a species extinction may be responsible for the observed association for this data. In addition, these islands are not a sample of any larger population of interest. Therefore, although the data are consistent with the theory that the odds of extinction decrease with increasing island size, the statistical analysis alone cannot be used to infer that the same relationship would exist in other islands or habitat areas or that the island size is a cause of extinction in these islands.

2. In 1846, the Donner party became stranded while crossing the Sierra Nevada mountains. See the file "5donner.txt" on blackboard for the data. we want to answer the question: Is there evidence that females have a greater chance of survival than for males, controlling for age?

   0. As a first step in using logistic regression, we must determine if extra-binomial variation exists. Does it?
   a. Write down (symbolically) the logistic model with logit link that would answer this question.
   b. Interpret the coefficients that you defined in 1a.
   c. Fit the model you defined in 1a. Include the output in your assignment.
   d. Find the estimated probability of survival for a 24 years-old male and a 30 years-old female.
   e. It is possible that there is an interaction between age and gender. Rewrite the model from 1a including this term.
   f. Interpret the coefficient for the interaction that you defined in 1e.
   g. Test to see if this term is warranted using the estimate divided by estimated standard error and the drop-in-deviance test. Is the drop-in-deviance test still applicable?
   h. Perform a goodness of fit test for the remaining model. *Note: there are several choices, some of which are not applicable to this situation. Choose one that is appropriate.*
   i Pretend that you are a consultant responding to a principal investigator's question. Write a short, convincing statement detailing your findings in this question. Include in your write up what you feel is the 'scope of inference' of this study. *Note: we haven't officially talked about this in class, but scope of inference refers to the broader conclusions that you may be able to draw from a study. Essentially, what, if anything, about this study is representative of a larger population of interest? For instance, can we use the conclusions from this study to make statements about the comparative survivability of men and women in: 1846, now, the U.S.A, etc.?* [2]

---

0. Extra-binomial variation isn't relevant to a study with no grouping.

[2] If you have gotten this far and haven't found that one of the subject's has a typo for gender, shame! Always remember to sanity check your data before using. Don't submit your assignment until you fix this.

a.

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{female} + \beta_2 \text{age}$$

where female is an indicator variable for gender = female.

b. β_0 is the log odds of survival of a male member of the party. $\beta_0 + \beta_1$ is the log odds of survival of a female. $\beta_1 = (\beta_0 + \beta_1) - \beta_0 =$ log odds ratio of a female to a male. β_2 corresponds to a change in log odds of survival with a change in age.

c.
```
> donner = read.table('5donner.txt',header=T)
> Y = donner$Survival
> Age = donner$Age
> Gender = donner$Sex
> out = glm(Y ~ Age + Gender,family=binomial(link=logit))
> print(summary(out))

Call:
glm(formula = Y ~ Age + Gender, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.23041    1.38686   2.329   0.0198 *
Age         -0.07820    0.03728  -2.097   0.0359 *
Gendermale  -1.59729    0.75547  -2.114   0.0345 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256

Number of Fisher Scoring iterations: 4
```

d. The log odds of surviving for each is -0.2438 and 0.8843, respectively. This corresponds to and estimated probability of survival of:

$$\frac{\exp\{-0.24377\}}{1+\exp\{-0.24377\}} = 0.4393$$

and

$$\frac{\exp\{0.8843\}}{1+\exp\{0.8843\}} = 0.7077125.$$

e.
```
Call:
glm(formula = Y ~ Age * Gender, family = binomial(link = logit))
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     7.24638    3.20517   2.261   0.0238 *
Age            -0.19407    0.08742  -2.220   0.0264 *
Gendermale     -6.92805    3.39887  -2.038   0.0415 *
Age:Gendermale  0.16160    0.09426   1.714   0.0865 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 47.346  on 41  degrees of freedom
AIC: 55.346
```

f. The change in log odds for males when Age changes is: $\beta_1 Age + \beta_3 Age$.

g. The first test is looking at the Wald test of comparing the estimate divided by it estimated standard error. This leads to a marginally significant p-value (0.0865). Alternatively, we can do a drop-in-deviance test (which is still valid since the difference in the deviances cancels out the saturated model comparison in each individual deviance), which is:

```
> 51.256 - 47.346
[1] 3.91
> pchisq(51.256 - 47.346,df=1,lower=F)
[1] 0.0479996
```

In general, the drop-in-deviance test is more accurate and should be preferred over the Wald test (since it is more approximately normal). In this case, the interaction is warranted for inclusion.

h. The overall deviance isn't approximately χ^2 as there is no grouping and therefore isn't useful in the situation. Here is some code:

```
donner = read.table('5donner.txt',header=T)
Y = donner$Survival
Age = donner$Age
Gender = donner$Sex
outInt = glm(Y ~ Age*Gender,family=binomial(link=logit))

fitVal = fitted(outInt)
ord    = order(fitVal,decreasing=FALSE)
n      = length(fitVal)
#I'm going to break it up into 9 groups
Observed = rep(0,9)
ResponseOrd = Y[ord]
fitValOrd   = fitVal[ord]
for(i in 0:8){
Observed[i+1] = sum(ResponseOrd[(4*i+1):(4*(i+1))] == 'yes')
}
avgProb = rep(0,9)
for(i in 0:8){
avgProb[i+1] = mean(fitValOrd[(4*i+1):(4*(i+1))])
}
```

```
Expected = avgProb*9
Hosmer = sum((Observed - Expected)^2/(9*avgProb*(1-avgProb)))
# Compare to ChiSq with 9 - 2 degrees of freedom
print(pchisq(Hosmer,df=9-2,lower=FALSE))
# Note, large values of Hosmer (and hence small pvals) indicate lack of fit
```

We get a HL of 15.899 with associated pval of 0.026. Note that this, along with a relatively high p-value for the Wald test (0.08) indicates that there is marginal evidence for including the interaction. Since interpretation is easier without it, it is probably best to leave it out of the model.

3. Returning to the Donner party data, let's redo part of the analysis with the probit link.

   a. What is the probit link?

   b. Fit the additive model of gender and age. What interpretation can you make of the estimated coefficients?

   d. Find the estimated probability of survival for a 24 years-old male and a 30 years-old female.

   e. Compare your answers briefly to the answers from the logistic output.

```
> ## Probit model
> donner = read.table('5donnerNoMistake.txt',header=T)
> Y = donner$Survival
> Age = donner$Age
> Gender = donner$Sex
> outProb = glm(Y ~ Age + Gender,family=binomial(link=probit))
> out = glm(Y ~ Age + Gender,family=binomial(link=logit))
> print(summary(out))

Call:
glm(formula = Y ~ Age + Gender, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.23041    1.38686   2.329   0.0198 *
Age         -0.07820    0.03728  -2.097   0.0359 *
Gendermale  -1.59729    0.75547  -2.114   0.0345 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
```

```
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256

Number of Fisher Scoring iterations: 4

> print(summary(outProb))

Call:
glm(formula = Y ~ Age + Gender, family = binomial(link = probit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7420  -1.0555  -0.2756   0.8861   2.0339

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.91730    0.76438   2.508   0.0121 *
Age         -0.04571    0.02076  -2.202   0.0277 *
Gendermale  -0.95828    0.43983  -2.179   0.0293 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.283  on 42  degrees of freedom
AIC: 57.283

Number of Fisher Scoring iterations: 5

> print(pnorm(1.9173 - 0.04571*24 - 0.95828))
[1] 0.4451123
> print(pnorm(1.9173 - 0.04571*30))
[1] 0.707467
```

4. For binary data, define a GLM using a log link. Show that effects refer to the relative risk. Why do you think this link is not often used (for binary data, we will see it again for Poisson regression)?

This corresponds to

$$\log(\pi(x)) = \beta_0 + \beta_1 x$$

which is equivalent to

$$\pi(x) = \exp\{\beta_0 + \beta_1 x\}.$$

We immediately see that since exp goes from 0 to ∞ and $\exp\{0\} = 1$, we need $\beta_0 + \beta_1 x < 0$ to get a valid probability. This limits its usefulness (though it doesn"t eliminate it. You can use this link function in SAS by setting link = log in proc genmod).

Also,

$$RR(x, x+1) = \frac{\pi(x+1)}{\pi(x)} = \exp\{\beta_1\}.$$