STAT 477/STAT 577
HW 7 - Solutions

Like many of their species, wolf spiders are known to practice cannibalism, with female spiders eating male spiders either before, during or after mating. However, since cannibalism does not occur after every act of mating, researchers have been interested in determining factors associated with the occurrence of cannibalism. In one such study, 52 female-male pairs were measured and then observed mating. Does the size difference between the female and male spiders help explain whether or not cannibalism occurred? The file **wolfspiders.csv** contains information about the presence or absence of cannibalism for each pair and the size difference between the female and male spiders (in mm).
To begin, read in the data file.

```
spiders.data<- read.csv(file.choose(), header = T)
```

Order the Cannibalism variable to make the category No the baseline category.

```
spiders.data$Cannibalism<- factor(spiders.data$Cannibalism,
                                  levels = c("No", "Yes"))
```

1. (2 pts) In what proportion of the 52 matings did cannibalism occur?

   There are several ways to find this proportion. I will use the `table()` function for the Cannibalism variable to find the proportion.

   ```
   table(spiders.data$Cannibalism)/52
   ```

   ```
   ##
   ##        No       Yes
   ## 0.7884615 0.2115385
   ```

   This value is 0.2115.

2. (4 pts) Write the equation for predicting the log odds of cannibalism from the size difference between the female and male spiders.

   Fit the logistic regression model for predicting cannibalism from the size difference.

   ```
   spiders.model<- glm(Cannibalism ~ Size.Diff, data = spiders.data,
                       family = binomial(link = "logit"))
   summary(spiders.model)
   ```

   ```
   ##
   ## Call:
   ## glm(formula = Cannibalism ~ Size.Diff, family = binomial(link = "logit"),
   ```

```
##     data = spiders.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0890     0.8288  -3.727 0.000194 ***
## Size.Diff     3.0693     1.0041   3.057 0.002237 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.663  on 51  degrees of freedom
## Residual deviance: 34.721  on 50  degrees of freedom
## AIC: 38.721
##
## Number of Fisher Scoring iterations: 6
```

Using the estimated intercept and slope from the output above, the equation for predicting the log odds of cannibalism from the size difference (x_i) is:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -3.0890 + 3.0693x_i$$

3. (4 pts) Write the equation for predicting the probability of cannibalism from the size difference between the female and male spiders.

   The equation for predicting the probability of cannibalism from the size difference (x_i) is:

$$\hat{p}_i = \frac{e^{-3.0890+3.0693x_i}}{1+e^{-3.0890+3.0693x_i}}$$

4. (10 pts; 5 each) Interpret the slope and intercept of the logistic regression equation.

   Slope: A 1mm increase in the size difference between the female and male spiders is associated with a $e^{3.0693} = 21.5268$ times increase in the predicted odds of cannibalism.

   Intercept: For a 0mm size difference between the female and male spiders (they are the same size), the predicted odds of cannibalism is $e^{-3.0890} = 0.0455$

5. (6 pts; 3 each) Find the predicted probability of cannibalism for a size difference between the female and male spiders of -0.2mm and 0.4mm.

   There are several ways to accomplish this task. I will enter the values of $x_i = -0.2$ and $x_i = 0.4$ into the equation in problem 3 and calculate the predicted probability.

   For $x_i = -0.2$:

$$\hat{p}_i = \frac{e^{-3.0890+3.0693(-0.2)}}{1+e^{-3.0890+3.0693(-0.2)}} = 0.02406$$

For $x_i = 0.4$:

$$\hat{p}_i = \frac{e^{-3.0890+3.0693(0.4)}}{1+e^{-3.0890+3.0693(0.4)}} = 0.1346$$

6. (16 pts; 8 each - 3 for interval and 5 for interpretation) Find confidence intervals for the probability of cannibalism for a size difference between the female and male spiders of 0mm and 0.8mm. Interpret both intervals.

   For the size difference of 0mm:

```
data1<- data.frame(Size.Diff = 0)
glm.prob.ci(spiders.model, newdata = data1, 0.95)

## [[1]]
##           2.5      97.5
## 1 0.008894204 0.1877545
```

   The CI is from 0.0089 to 0.1878. This means we have 95% confidence the probability of cannibalism in the population in matings that occur between the female and male spiders of the same size is between 0.0089 and 0.1878.

   For the size difference of 0.8mm:

```
data2<- data.frame(Size.Diff = 0.8)
glm.prob.ci(spiders.model, newdata = data2, 0.95)

## [[1]]
##         2.5      97.5
## 1 0.1831188 0.5567838
```

   The CI is from 0.1831 to 0.5568. This means we have 95% confidence the probability of cannibalism in the population in matings that occur between a female spider 0.8mm larger than the male spider is between 0.1831 and 0.5568.

7. (20 pts) Test for the statistical significance of the size difference between the female and male spiders in predicting the probability of cannibalism. Report both the Wald test statistic and the likelihood ratio test statistic as a part of your answer.

   The hypothesis test will have null and alternative hypotheses of:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

   The Wald test statistic comes from Coefficients table in the summary output of the model. The test statistic $z = 3.057$ with p-value $= 0.0022$. This means we have very strong evidence the size difference in the female and male spiders is associated with the probability of cannibalism in this population.

   The Likelihood test statistic can be found using the `anova()` function.

```
anova(spiders.model, test = "Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Cannibalism
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         51     53.663
## Size.Diff  1   18.942        50     34.721 1.348e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Likelihood ratio test statistic is 18.942 with a p-value < 0.0001. This means we have extremely strong evidence the size difference in the female and male spiders is associated with the probability of cannibalism in this population.

8. (3 pts) Calculate the pseudo R^2 statistic for this logistic regression. Comment on its value.

   We can calculate this value using the `McFR2()` function

```
McFR2(spiders.model)

## [1] 0.3529793
```

   A value of 0.3530 indicates a good model.

9. (12 pts) Conduct a goodness of fit test using the Hosmer-Lemeshow test statistic with the number of groups set to 5. Does this model appear to fit the data?

   First, we will need to credit a new variable consisting of 1s and 0s for the Cannibalism variable

```
spiders.data$Cannibalism.1.0<- ifelse(spiders.data$Cannibalism == "Yes", 1, 0)
```

   Now we can run the Hosmer-Lemeshow test.

```
hoslem.test(spiders.data$Cannibalism.1.0,
            spiders.model$fitted.values, g = 5)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  spiders.data$Cannibalism.1.0, spiders.model$fitted.values
## X-squared = 3.0035, df = 3, p-value = 0.3911
```

Null hypothesis: model is a good fit for the data
Alternative hypothesis: model is not a good fit for the data

The test statistic is $X^2 = 3.0035$ with a p-value = 0.3911. We conclude we have no evidence of lack of model fit.

10. (11 pts) Give the confusion table for the logistic regression model. Use this table to calculate the agreement, sensitivity, and specificity of the model. Comment on these values.

    We will use the `confusion.glm()` function.

```
confusion.glm(spiders.model)
```

```
## $`Confusion Table`
##          predicted
## observed  0  1
##        0 39  2
##        1  8  3
##
## $Agreement
## [1] 0.8076923
##
## $Sensitivity
##         1
## 0.2727273
##
## $Specificity
##         0
## 0.9512195
```

    The figures indicate we correctly predicted 80.77% (Agreement) of the 52 observations. However, while we correctly predicted 95.12% (Specificity) cases when Cannibalism did not occur, we only correctly predicted 27.27% (Sensitivity) cases when Cannibalism did occur.

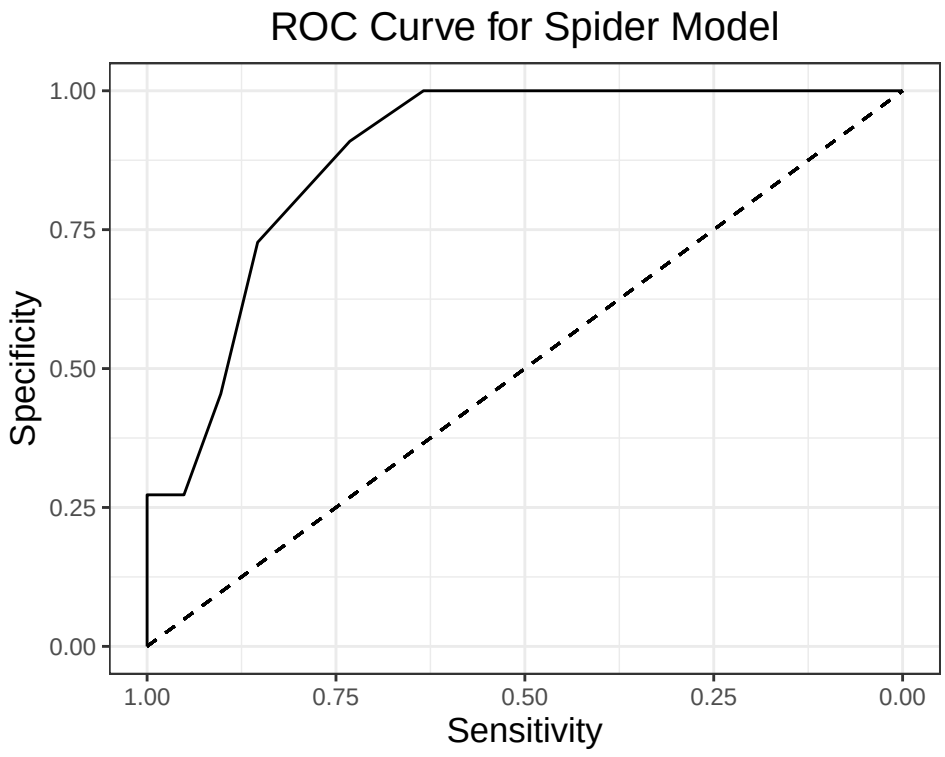
11. (12 pts) Graph the ROC curve for this logistic regression model. Calculate the area under the ROC curve and interpret this value.

    The ROC curve is:

```
spiders.roc<- roc(spiders.data$Cannibalism ~ spiders.model$fitted.values)

## Setting levels:  control = No, case = Yes
## Setting direction:  controls < cases

ggroc(spiders.roc)+
  theme_bw()+
  theme(axis.title.y = element_text(size = rel(1.2)))+
  theme(axis.title.x = element_text(size = rel(1.2)))+
  theme(axis.text.x = element_text(size = rel(1)))+
  theme(axis.text.y = element_text(size = rel(1)))+
  theme(plot.title = element_text(hjust=0.5, size = rel(1.4)))+
  geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
                linetype="dashed")+
  labs(x = "Sensitivity",
       y = "Specificity",
       title = "ROC Curve for Spider Model")
```



The area under this curve is:

```
auc(spiders.roc)

## Area under the curve: 0.8869
```

This means that if I randomly select a case where Cannibalism occurred (success) and randomly select a case where Cannibalism did not occur (failure), the probability the success will have a higher predicted probability than the failure is 0.8869.