

## Exercise 8

### ST2304 Exercises Week 13: Binomial Models

#### Solutions

##### Question 1

Fit the model with linear and quadratic terms, as above. Use `anova()` to compare the models (or use AIC if you want to try that!)

1. Would you include the precipitation and temperature in the model?
2. Would you include quadratic terms for either or both of these models?

We will do backwards model selection starting with a full model with precipitation and temperature and their quadratic terms:

```
# Start with the full model
m1<-glm(cbind(NPres, NAbs) ~ prec.mean.sc + temp.mean.sc + I(prec.mean.sc^2) +
        I(temp.mean.sc^2), family = "binomial", data = DryoPil)
summary(full)
```

We can already see from the model summary that temperature and its quadratic term are not significant.

```
## start reducing full model using anova()

m1.1<-glm(cbind(NPres, NAbs) ~ temp.mean.sc + I(prec.mean.sc^2) +
          I(temp.mean.sc^2), family = "binomial", data = DryoPil)

m1.2<-glm(cbind(NPres, NAbs) ~ prec.mean.sc + I(prec.mean.sc^2) +
          I(temp.mean.sc^2), family = "binomial", data = DryoPil)

m1.3<-glm(cbind(NPres, NAbs) ~ prec.mean.sc + temp.mean.sc +
          I(temp.mean.sc^2), family = "binomial", data = DryoPil)

m1.4<-glm(cbind(NPres, NAbs) ~ prec.mean.sc + temp.mean.sc + I(prec.mean.sc^2),
          family = "binomial", data = DryoPil)

anova(m1.1,m1,test="Chisq")
anova(m1.2,m1,test="Chisq")
anova(m1.3,m1,test="Chisq")
anova(m1.4,m1,test="Chisq")
```

The output shows a significant p-value for all model comparisons except between models `m1` and `m1.4`. A significant p-value means that the term we are trying to remove is significant and needs to be kept in the model.

So we use model m1.4 as a reduced form of the full model and continue reducing it, one term at a time.

The summary of m1.4 shows that the variable "temp.mean.sc" is not significant.

```
m1.4.1<-glm(cbind(NPres, NAbs) ~ temp.mean.sc + I(prec.mean.sc^2),
            family = "binomial", data = DryoPil)
m1.4.2<-glm(cbind(NPres, NAbs) ~ prec.mean.sc + I(prec.mean.sc^2),
            family = "binomial", data = DryoPil)
m1.4.3<-glm(cbind(NPres, NAbs) ~ prec.mean.sc + temp.mean.sc ,
            family = "binomial", data = DryoPil)

anova(m1.4.1,m1.4,test="Chisq")
anova(m1.4.2,m1.4,test="Chisq")
anova(m1.4.3,m1.4,test="Chisq")
```

The output shows a significant p-value for all model comparisons except between m1.4 and m1.4.2. So we continue reducing model m1.4.2, one term at a time.

```
m1.4.2.1<-glm(cbind(NPres, NAbs) ~ I(prec.mean.sc^2),
              family = "binomial", data = DryoPil)
m1.4.2.2<-glm(cbind(NPres, NAbs) ~ prec.mean.sc ,
              family = "binomial", data = DryoPil)

anova(m1.4.2,m1.4.2.1,test="Chisq")
anova(m1.4.2,m1.4.2.2,test="Chisq")
```

Model m1.4.2 cannot be reduced any further and is our best and final model. Thus, we include only precipitation and its quadratic term in the model.

3. Look at the parameter estimates, using summary(). Qualitatively, are there any differences between the models? Obviously the parameters will be different, but are there any large differences, and can you explain why? (if you can't work this out from just the summary, try plotting the predictions).

Comparing summaries of models m1, m1.4 and m1.4.2, we see that the parameter estimates do not change drastically, probably because the variables that have a significant effect (prec.mean.sc + I(prec.mean.sc^2)) are included in these models.

4. Is there any evidence of over-dispersion in the model? If there is, does it change the model you would chose, and if so, how?

```
pchisq(deviance(m1.4.2), df.residual(m1.4.2), lower.tail = FALSE) ##p-value< 0.05
Disp <- deviance(m1.4.2)/df.residual(m1.4.2)
Disp ## =1.78437
```

The low p-value in the goodness of fit test suggests that the data does not fit the model well. This is because of overdispersion in the model (estimate of overdispersion is 1.78). What we can do is to correct for this overdispersion by increasing the standard errors by  $\sqrt{1.78}$  times.

## Problem 2: Bumpus' Sparrows

Fit a model with Total Length ("Total.Length") and Sex as covariates (just as main effects), and survival as a response.

1. What effect does the total length have: do larger or smaller birds survive more?

```
Call:
glm(formula = survived ~ sex + Total.Length, family = "binomial",
     data = Bumpus)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7676	-1.1132	0.6234	0.9831	1.8951

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	33.27342	9.32686	3.567	0.000360	***
SexMale	1.23911	0.42797	2.895	0.003788	**
Total.Length	-0.21273	0.05917	-3.595	0.000324	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Increase in length decreases the probability of survival. So smaller birds survive more/better.

2. What is the difference in the odds of survival of birds that are 5mm different in length?

With the logit link you have:  $\log(p/(1-p)) = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{length}$

The odds of survival is thus:  $p/(1-p) = \exp(\beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{length})$

So changing length by 5mm changes the odds by a factor of  $\exp(-0.21273 \cdot 5) = 0.34$ . This factor is a ratio between two odds, hence the term odds-ratio for this. Since the effects on the odds are multiplicative, we can say that the odds of survival decreases by  $1 - 0.34 = 0.66 = 66\%$  when increasing length by 5mm.

3. Does Sex have an effect on survival?

Sex has a significant effect on survival, with males surviving better than females.

Add an interaction between Sex and Total Length.

4. What is the effect of the interaction? Is the estimated effect of body length stronger or weaker in males (i.e. is the slope of the effect steeper)?

Call:

```
glm(formula = survived ~ Sex * Total.Length, family = "binomial",  
     data = Bumpus)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0301	-1.0384	0.4452	1.0775	1.8204

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	12.56849	12.89591	0.975	0.3298
SexMale	41.92464	19.46386	2.154	0.0312 *
Total.Length	-0.08142	0.08169	-0.997	0.3189
SexMale:Total.Length	-0.25556	0.12201	-2.095	0.0362 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The effect of total length is even more negative for males (goes from an estimated effect of -0.08 to -0.25). Thus, the negative effect of total length on survival is even stronger for males, compared to females.

5. Use analysis of deviance (i.e. anova()) to compare the models with and without an interaction. Would you include an interaction?

```
## Try reducing the full model, hence the interaction term  
mod1<-glm(Survived ~ Sex*Total.Length,  
          family = "binomial", data = Bumpus)  
  
mod1.1<-glm(Survived ~ Sex+Total.Length,  
            family = "binomial", data = Bumpus)  
  
anova(mod1.1,mod1,test="Chisq")##p-value< 0.05
```

A significant p-value means that the interaction term is significant and cannot be reduced. We therefore need to keep it in the model.