MPulleyM2Q4

Melissa Pulley

2024-04-23

```
cod = read.csv("parasitecod.csv")
```

1. Before running the model, please note that both year and region (area in the dataset) have numeric values. However, we want them to be categorical variables. Transform the dataset so that they become categorical.

```
cod$Year = as.factor(cod$Year)
cod$Area = as.factor(cod$Area)

code <- subset(cod, select = c(Length, Prevalence, Year, Area))

summary(code)
```

```
##      Length         Prevalence       Year     Area
##  Min.   : 17.00   Min.   :0.0000   1999:567   1:272
##  1st Qu.: 44.00   1st Qu.:0.0000   2000:230   2:255
##  Median : 54.00   Median :0.0000   2001:457   3:415
##  Mean   : 53.45   Mean   :0.4785              4:312
##  3rd Qu.: 62.00   3rd Qu.:1.0000
##  Max.   :101.00   Max.   :1.0000
##  NA's   :6
```

2. Run the model

```
m = glm(Prevalence~Length+Year+Area, family = binomial(link = "logit"),data=code)
summary(m)
```

```
##
## Call:
## glm(formula = Prevalence ~ Length + Year + Area, family = binomial(link = "logit"),
##     data = code)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.465947   0.269333  -1.730 0.083629 .
## Length       0.009654   0.004468   2.161 0.030705 *
## Year2000     0.566536   0.169715   3.338 0.000843 ***
## Year2001    -0.680315   0.140175  -4.853 1.21e-06 ***
## Area2       -0.626192   0.186617  -3.355 0.000792 ***
## Area3       -0.510470   0.163396  -3.124 0.001783 **
## Area4        1.233878   0.184652   6.682 2.35e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1727.8  on 1247  degrees of freedom
## Residual deviance: 1537.6  on 1241  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 1551.6
##
## Number of Fisher Scoring iterations: 4
```

3. Describe the results. Which predictors are significant?

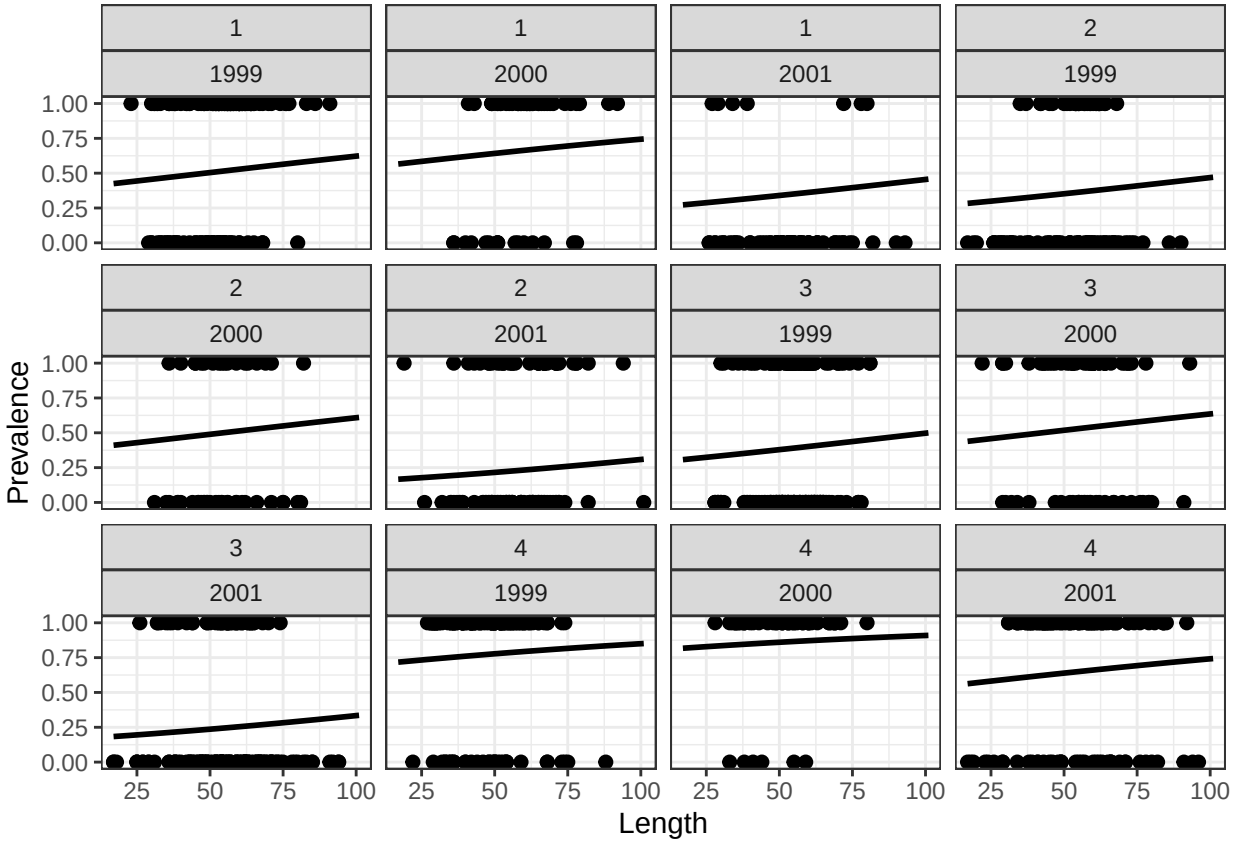
   The results give the coefficients for the glm model with link function "logit". Almost all of the predictors have significant coefficients. The only one not significant is the intercept, which means we do not have function evidence to support a nonzero intercept.

4. Now, let's plot the model. The Y axis should be prevalence (between 0 and 1) and you should include ALL explanatory variables in the plot. Tip: look at the logistic lab (classlogistic) and how we plotted the 3 habitats in that (using facet_wrap).

```
newdata2 = NaN
newdata2 <- expand.grid(Length = seq(min(code$Length,na.rm = TRUE), max(code$Length,na.rm=TRUE)),
                        Year = c("1999", "2000", "2001"),
                        Area = as.factor(seq(1,4)))
#pred = m$coefficients[1] + sum(m$coefficients[2:7])*newdata2$Length

pred.link.full <- predict(m, newdata = newdata2, se.fit = TRUE)

newdata2$p <- plogis(pred.link.full$fit) # back transform to probability scale
newdata2$lower <- plogis(pred.link.full$fit - 1.96 * pred.link.full$se.fit)
newdata2$upper <- plogis(pred.link.full$fit + 1.96 * pred.link.full$se.fit)

summary(newdata2)
```

```
##      Length      Year     Area          p              lower
##  Min.   : 17   1999:340   1:255   Min.   :0.1668   Min.   :0.1112
##  1st Qu.: 38   2000:340   2:255   1st Qu.:0.3456   1st Qu.:0.2798
##  Median : 59   2001:340   3:255   Median :0.4941   Median :0.4095
##  Mean   : 59              4:255   Mean   :0.5145   Mean   :0.4362
##  3rd Qu.: 80                      3rd Qu.:0.6597   3rd Qu.:0.5834
##  Max.   :101                      Max.   :0.9097   Max.   :0.8492
##      upper
##  Min.   :0.2426
##  1st Qu.:0.4220
##  Median :0.5837
##  Mean   :0.5917
##  3rd Qu.:0.7339
##  Max.   :0.9474
```

```
#predict= m$coefficients[1] + sum(m$coefficients[2:6])*newdata2

library(ggplot2)
ggplot() +
 geom_point(data = code, aes(x = Length, y = Prevalence),size=2) +
  geom_path(data=newdata2, aes(x = Length, y = p),size=1) +
  xlab("Length") +
  ylab("Prevalence") +
  facet_wrap(~Area+Year)+
  theme_bw()
```


1
1999
1
2000
1
2001
2
1999
2
2000
2
2001
3
1999
3
2000
3
2001
4
1999
4
2000
4
2001
Prevalence
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Length