453 HW4

Max Tjen

2023-07-27

Packages

```
library(nnet)
library(dplyr)
library(tidyverse)
```

Question 10

```
data10 <- read_csv("/Users/mtjen/Desktop/453/hw5/DeathPenalty.csv") |>
  mutate(STATUS = factor(STATUS),
         LEGAL = factor(LEGAL),
         CONFLICT = factor(CONFLICT))
```

Part A

```
# build model
mod10 <- multinom(STATUS ~ LEGAL + HDI + GINI + GNI +
                    LITERACY + URBAN + POL + CONFLICT,
                  data = data10,
                  trace = FALSE)

round(AIC(mod10), 3)
```

```
## [1] 311.809
```

```
round(BIC(mod10), 3)
```

```
## [1] 461.833
```

```
# get values for corrected AIC
extAIC <- extractAIC(mod10)
modDf <- extAIC[1]
aicVal <- extAIC[2]
```

```
numRows = nrow(data10)

aicC <- aicVal + ((2 * modDf * (modDf + 1)) / (numRows - modDf - 1))
round(aicC, 3)
```

[1] 372.082

The model's AIC is 311.8, BIC is 461.8, and AICc is 472.1.

Part B

Status can be viewed as an ordinal variable because there is an apparent order in death penalty severity,
with status a (public execution) being the most extreme and status d (no death penalty) being the least
severe. Overall, the order is status a is more severe than status b, status b is more severe than status c, and
status c is more severe than status d.

Part C

```
# make status ordinal
data10c <- data10 |>
  mutate(STATUS = factor(STATUS, order = TRUE,
                         levels = c("a", "b", "c", "d")))

# build model
mod10c <- MASS::polr(STATUS ~ LEGAL + HDI + GINI + GNI +
                       LITERACY + URBAN + POL + CONFLICT,
                     data = data10, method = "logistic")

round(AIC(mod10c), 3)
```

[1] 296.911

```
round(BIC(mod10c), 3)
```

[1] 352.803

```
# get values for corrected AIC
extAIC <- extractAIC(mod10c)
modDf <- extAIC[1]
aicVal <- extAIC[2]
numRows = nrow(data10c)

aicC <- aicVal + ((2 * modDf * (modDf + 1)) / (numRows - modDf - 1))
round(aicC, 3)
```

[1] 303.245

The model's AIC is 296.9, BIC is 352.8, and AICc is 303.2.

Part D

There is not evidence that these ordinality assumptions lead to poor model fit, as AIC, BIC, and corrected AIC values are better for the proportional odds model than the multinomial model.

Question 23

```
data23 <- read_csv("/Users/mtjen/Desktop/453/hw5/HorseshoeCrabs.csv")
```

Part A

```
mod23 <- glm(Sat ~ Color + Spine + Width + Weight,
             family = poisson(link = "log"),
             data = data23)

car::Anova(mod23)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Sat
##         LR Chisq Df Pr(>Chisq)
## Color     7.9690  1   0.004759 **
## Spine     0.5009  1   0.479111
## Width     0.3221  1   0.570374
## Weight    8.3329  1   0.003893 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By running an ANOVA test on the poisson regression model, we can see that a female crab's color and weight are statistically significant in predicting the number of satellite males in her vicinity. This signals that there are associations between female color and satellite males as well as female weight and satellite males.

Part B

```
dev <- mod23$deviance
df <- mod23$df.residual

dev/df
```
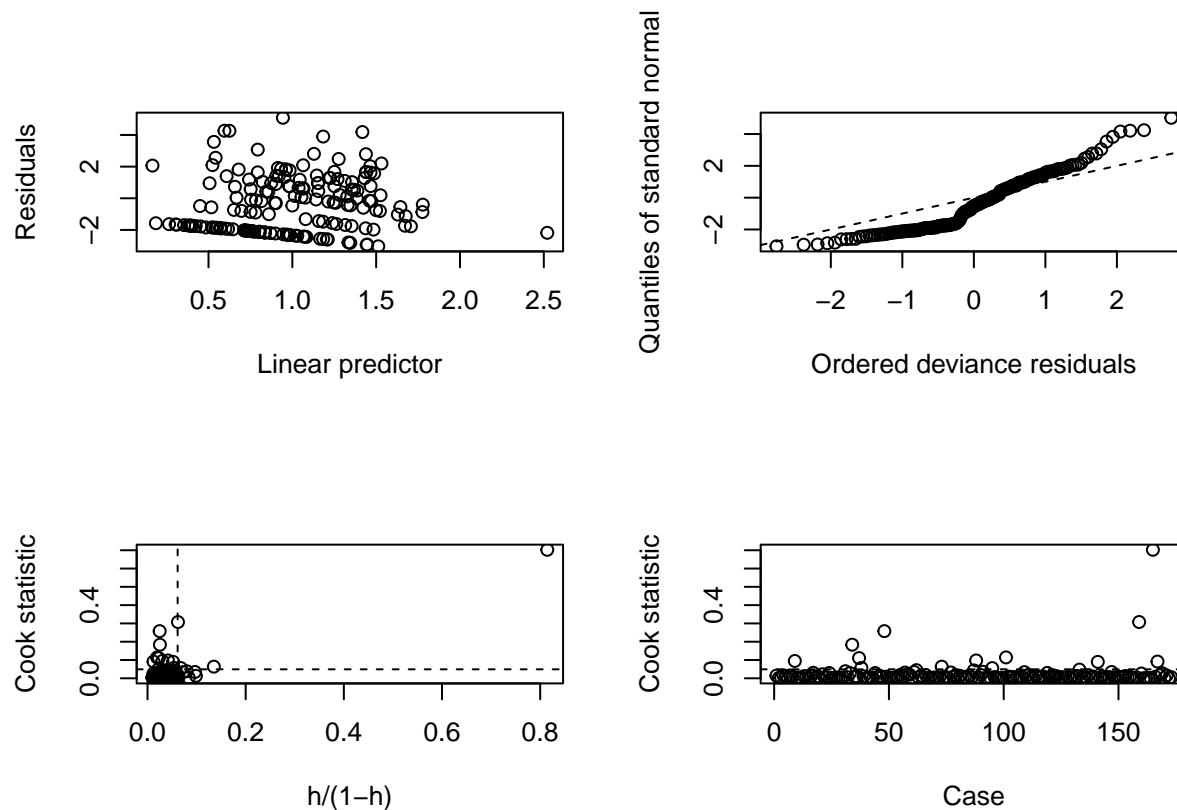
```
## [1] 3.284703
```

The model's deviance/df value is 3.285, which is much higher than 1. This indicates that the model is over dispersed and doesn't fit well.

Part C

```
boot::glm.diag.plots(mod23)
```


From these residual plots, it is looks like there may be a couple influential outlier points in the data, with one being particularly influential. The residuals also don't look very normal, which may potentially indicate the model not fitting very well.

Part D

```
PostFitGOFTest = function(obs, pred, g = 0) {
  if(g == 0) g = round(min(length(obs)/5,20))
 ord <- order(pred)
 obs.o <- obs[ord]
 pred.o <- pred[ord]
 interval = cut(pred.o, quantile(pred.o, 0:g/g), include.lowest = TRUE)  # Creates factor with levels 1
 counts = xtabs(formula = cbind(obs.o, pred.o) ~ interval)
 centers <- aggregate(formula = pred.o ~ interval, FUN = "mean")
 pear.res <- rep(NA,g)
 for(gg in (1:g)) pear.res[gg] <- (counts[gg] - counts[g+gg])/sqrt(counts[g+gg])
 pearson <- sum(pear.res^2)
 if (any(counts[((g+1):(2*g))] < 5))
  warning("Some expected counts are less than 5. Use smaller number of groups")
 P = 1 - pchisq(pearson, g - 2)
 cat("Post-Fit Goodness-of-Fit test with", g, "bins", "\n", "Pearson Stat = ", pearson, "\n", "p = ", P
```

```
return(list(pearson = pearson, pval = P, centers = centers$pred.o, observed = counts[1:g], expected = (
}
```

```
observed <- data23$Sat
predicted <- round(predict(mod23,
                           newdata = data23 |> select(-Sat),
                           type = "response"), 0)
```

```
#PostFitGOFTest(observed, predicted)
```

provided function returns an error:

Question 25

```
# identify crab with outlier weight
data25 <- data23 |>
  filter(Weight != 5.20)
```

```
nrow(data23)
```

```
## [1] 173
```

```
nrow(data25)
```

```
## [1] 172
```

Part A

```
mod25 <- glm(Sat ~ Color + Spine + Width + Weight,
             family = poisson(link = "log"),
             data = data25)
```

```
car::Anova(mod25)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Sat
##          LR Chisq Df Pr(>Chisq)
## Color      6.8888  1  0.0086736 **
## Spine      0.2247  1  0.6354770
## Width      0.0028  1  0.9574351
## Weight    12.7318  1  0.0003595 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By running an ANOVA test on the poisson regression model, we can see that a female crab's color and weight are again statistically significant in predicting the number of satellite males in her vicinity. This signals that there are associations between female color and satellite males as well as female weight and satellite males.

Part B

```
dev <- mod25$deviance
df <- mod25$df.residual

dev/df
```
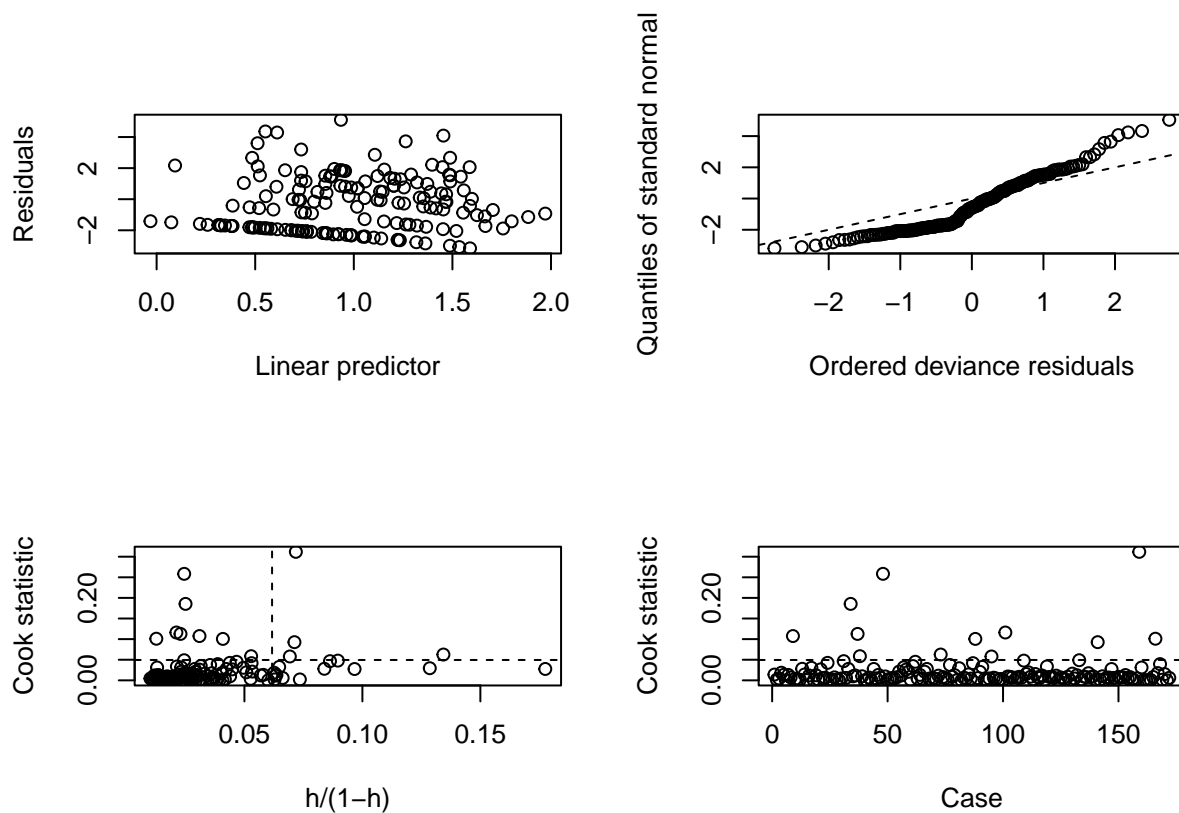
```
## [1] 3.27596
```

The model's deviance/df value is 3.276, which is still much higher than 1. This indicates that the model is over dispersed and doesn't fit well.

Part C

```
boot::glm.diag.plots(mod25)
```


As before, it is looks like there may still be a couple influential outlier points in the data. The residuals also don't look very normal, which may potentially indicate the model not fitting very well.