Homework 2: Logistic Regression

STAT 340: Applied Regression

Nonconstant error variance in the linear-probability model

From your reading, you will note that one of the problems with the linear-probability model is that the equal variance assumption that we need for a linear regression model is violated. Explore this violation in the following exercise:

Exercise 14.1 (J. Fox, 3rd Edition): Make a table showing the variance of the error $Var(\epsilon) = \pi(1-\pi)$ for the following values of π: 0.001, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 0.999. For what values of π is the heteroskedasticity (unequal variance) problem serious?

Model Specification

For the rest of this homework assignment, we will continue using the crab species identification data set from the class notes.

```
## Sample data set and make reproducible
set.seed(478)
crabs <- MASS::crabs[sample.int(nrow(crabs)),]
crabs <- crabs %>% dplyr::select(-index)

## Add 0/1 response variable
crabs <- crabs %>%
  mutate(
    sp_01 = ifelse(sp == "O", 1, 0)
  )
head(crabs)
```

```
##   sp sex   FL   RW   CL   CW   BD sp_01
## 1  O   F 21.4 18.0 41.2 46.2 18.7     1
## 2  O   M 15.1 11.4 30.2 33.3 14.0     1
## 3  O   M 18.8 13.4 37.2 41.1 17.5     1
## 4  O   F 22.5 17.2 43.0 48.7 19.8     1
## 5  O   M 14.2 10.7 27.8 30.9 12.7     1
## 6  B   M 17.9 14.1 39.7 44.6 16.8     0
```

```
## Model
crabs_logit <- glm(sp_01 ~ sex*FL + sex*CL, data=crabs, family=binomial)
summary(crabs_logit)
```

```
##
## Call:
## glm(formula = sp_01 ~ sex * FL + sex * CL, family = binomial,
```

```
##     data = crabs)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.19438  -0.07286   0.00001   0.01700   2.45670
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -12.073      4.671  -2.585 0.009749 **
## sexM          -7.227      8.420  -0.858 0.390775
## FL             9.922      2.771   3.581 0.000343 ***
## CL            -4.434      1.252  -3.542 0.000397 ***
## sexM:FL        9.486      7.211   1.316 0.188333
## sexM:CL       -4.249      3.236  -1.313 0.189104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 277.259  on 199  degrees of freedom
## Residual deviance:  35.442  on 194  degrees of freedom
## AIC: 47.442
##
## Number of Fisher Scoring iterations: 9
```

Write out the model associated with ... Be sure to include the three components of the GLM, and define any notation (e.g., β, X) that you use.

Write down the design matrix associated with the linear component of the GLM from the previous question.

Check assumptions

Check the binary response, linearity, and independence assumptions for the model specified in the previous problem (for now we are going to disregard the variance assumption). In particular, make sure to check linearity on the appropriate scale - your explanatory variable of interest should be on the x-axis, and the empirical (calculated from the data) log odds of the response should be on the y-axis. To do this, you will need to make a new data frame with $\log\left(\frac{\pi}{1-\pi_i}\right)$ in one column, and the explanatory variables in the other column.

Suppose linearity is not satisfied for a logistic regression model. What would you consider doing to remedy this assumption violation? (Hint, think about what you have done in other models to address departures from linearity. You may give a general answer.)

Hypothesis tests

Is there evidence that the effect of CL or FL on the odds of being an orange crab differs for male crabs (versus female crabs)? Write out the appropriate hypotheses, and carry out the hypothesis test.

Compare the model fit (`crabs_logit`) to a model without the interactions with sex using AIC. What model would you favor based on the AIC value? Why?