Logistic regression with R

The data for this exercise consist of discrete outcomes: None, Some, or Marked improvement following treatment for rheumatoid arthritis. Patients are classified by gender, age, and whether they received an active treatment or placebo. How can we construct a model to predict or explain the outcome of improvement? We will start with a binary response variable (logistic regression), and then extend this to the three-level response (proportional odds model). A script for this exercise is in `logistic-arthritis.R`.

In R, the data are available as the data frame `Arthritis` in the **vcd** package. For logistic regression, we need to explicitly create the binary variable `Better` from the 3-level response `Improved`.

```
data(Arthritis, package="vcd")
str(Arthritis)
Arthritis$Better <- Arthritis$Improved > "None"
```

Note that `Treatment` and Sex are factors. The binary outcome, `Better`, is defined corresponding to the distinction of `None` vs. (`Some, Marked`).

1. Carry out a simple linear regression predicting better from age. We'll use the visreg package to visualize the fitted model.

   ```
   arth.lm <- lm(Better ~ Age, data=Arthritis)
   summary(arth.lm)

   # plot the data and the fitted line
   library(visreg)
   visreg(arth.lm, points.par=list(cex=1.2))
   ```

   What are some possible problems with this model? [Think about assumptions.]

2. Now, try a logistic regression model.

   ```
   arth.logistic0 <- glm( Better ~ Age, data=Arthritis, family=binomial)
   summary(arth.logistic0)
   visreg(arth.logistic0, scale="response", jitter=TRUE)
   ```

3. Let's add main effects of sex and treatment to the model.

   ```
   arth.logistic <- glm( Better ~ Age + Sex + Treatment,
           data=Arthritis, family=binomial)
   summary(arth.logistic)
   visreg(arth.logistic, scale="response")
   ```

Examine the plots to interpret the effects of age, sex and treatment on the probability of improvement.

4. As a screening procedure for more complex models, we test a model that includes all two-way terms, to ask if there are any higher-order terms worth considering. What do you conclude?

```
# fit model with all 2-way interactions
arth.logistic2 <- glm( Better ~ (Age + Sex + Treatment)^2, data=Arthritis,
family=binomial)
# test individual terms
Anova(arth.logistic2)
```

5. We can compare nested models using the `anova()` function. This gives likelihood ratio tests.

```
# compare models
anova(arth.logistic0, arth.logistic, arth.logistic2, test="Chisq")
```

6. As a final step, we shift from trying to model the binary outcome, `better`, to a model for the polytomous response, `Improved`, whose values are to None, Some, or Marked improvement. The simplest model is the **proportional odds model**, which requires equal slopes for the adjacent-category logits. In R, this uses the function polr() in the MASS library.

```
library(MASS)
arth.polr <- polr( Improved ~ Age + Sex + Treatment, data=Arthritis)
summary(arth.polr)
```