Practice 10

You need to use the R packages faraway and MASS to work on the following questions.

The nes96 dataset is a 10-variable subset of the 1996 American National Election Study. Missing values and "don't know" responses have been listwise deleted. Respondents expressing a voting preference other than Clinton or Dole have been removed. As the result, nes96 contains 944 observations on the following 10 variables:

- popul: population of respondent's location in 1000s of people
- TVnews: days in the past week spent watching news on TV
- selfLR: Left-Right self-placement of respondent: an ordered factor with levels extremely liberal, extLib < liberal, Lib < slightly liberal, sliLib < moderate, Mod < slightly conservative, sliCon < conservative, Con < extremely conservative, extCon
- ClinLR: Left-Right placement of Bill Clinton (same scale as selfLR): an ordered factor with levels extLib < Lib < sliLib < Mod < sliCon < Con < extCon
- DoleLR: Left-Right placement of Bob Dole (same scale as selfLR): an ordered factor with levels extLib < Lib < sliLib < Mod < sliCon < Con < extCon
- PID: Party identification: an ordered factor with levels strong Democrat, strDem < weak Democrat, weakDem < independent Democrat, indDem < independent independent, indind < indepedent Republican, indRep < waek Republican, weakRep < strong Republican, strRep
- age: Respondent's age in years
- educ: Respondent's education: an ordered factor with levels 8 years or less, MS < high school dropout, HSdrop < high school diploma or GED, HS < some College, Coll < Community or junior College degree, CCdeg < BA degree, BAdeg < postgraduate degree, MAdeg
- income: Respondent's family income: an ordered factor with levels

  $3Kminus < $3K-$5K < $5K-$7K < $7K-$9K < $9K-$10K < $10K-$11K < $11K-$12K < $12K-$13K < $13K-$14K < $14K-$15K < $15K-$17K < $17K-$20K < $20K-$22K < $22K-$25K < $25K-$30K < $30K-$35K < $35K-$40K < $40K-$45K < $45K-$50K <

  $50K-$60K < $60K-$75K < $75K-$90K < $90K-$105K < $105Kplus
- vote: Expected vote in 1996 presidential election: a factor with levels Clinton and Dole

1. Type help(nes96) to see its description. Conduct an exploratory data analysis on nes96 to better understand the nes96 data. For example, check the size of the data, the type of each variable (categorical, factor, ordered factor, numerical), etc.

2. For simplicity, we consider only the age, education level and income group of the respondents. The response variable of our interest will be the party identification (`PID`) of the respondent which has 3 levels: `Democrat`, `Independent` and `Republican`, and is of ordinal nature if `Independent` can be regarded as somewhere in between `Democrat` and `Republican` in regard to political view. The original data involved more than three categories for `PID`; so again for simplicity of the presentation we collapse this to three, which will be saved in variable `party`. Moreover, we over-write the `income` factor by an `income` score variable. The following `R` commands implement the above discussions, and create a new data.frame `rnes96` to include `age`, `education`, `income` and `party` variables:

   ```
   party <- nes96$PID
   levels(party) <- c("Democrat","Democrat","Independent","Independent","Independent","Republican","Republican")
   inca <- c(1.5,4,6,8,9.5,10.5,11.5,12.5,13.5,14.5,16,18.5,21,23.5, 27.5,32.5,37.5,42.5,47.5,55,67.5,82.5,97.5,115)
   income <- inca[unclass(nes96$income)]
   table(nes96$income)
   table(income)
   rnes96 <- data.frame(party, income, education=nes96$educ, age=nes96$age)
   summary(rnes96)
   ```

   Conduct an exporatory data analysis on `rnes96` to make sure you are clear about its correspondence with the original data `nes96`.

3. Use the `polr` function in `MASS` to fit a proportional odds model on `party` with `age`, `education` and `income` as the predictors including their main effects only. Save the results into `pomod`. Then explore `pomod` using the commands such as `summary`, `anova`,`fitted`, `prediction`, `deviance` and `resid` etc. to see whether you understand the `R` outcomes and are able to interpret them. You may need use the following `R` commands:

   ```
   pomod <- polr(party~age+education+income, data=rnes96, Hess=T,method="logistic")
   pomod
   summary(pomod)
   anova(pomod)

   fitted(pomod)
   pomod$fitted
   resid(pomod)  #does not exist.
   pomod$residual
   deviance(pomod)
   pomod$deviance
   pomod$lp
   predict(pomod,type="probs")
   predict(pomod,type="class")
   table(predict(pomod,type="class"))
   ```

4. In `pomod`, predictor `education` is used as an ordinal factor. Now refit the model by using education as a nominal factor (denoted as `educ.f`). Save the result into `pomodf`. Then compare `pomodf` with `pomod`. You may need use the following `R` commands:

   ```
   rnes96$educ.f <-factor(unclass(rnes96$education))
   pomodf <- polr(party~age+educ.f+income, data=rnes96, Hess=T,method="logistic")
   summary(pomodf)
   anova(pomod, pomodf)
   ```

5. Perform variable selection based on `pomod` using `step` function with AIC or BIC option. Save the results into `pomod.aic` and `pomod.bic` respectively.

6. Compare `pomod` with `pomod.aic` and `pomod.bic`. Then find the best model among the three.

7. Explain the meaning of the estimated coefficient of `income` in the best model.

8. For the last respondent in the data, compute the predicted probabilities of the person's 3 possible party membership based on the best model among `pomod`, `pomod.aic` and `pomod.bic`. Then find approximately 95% confidence intervals for the three probabilities.