Practice 10 Solutions

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 10 variables.

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.

```
library(faraway); data(nes96); help(nes96); dim(nes96) #n=944, p=10
library(MASS); head(nes96); summary(nes96)
is.ordered(nes96$income); levels(nes96$income); levels(nes96$PID); is.ordered(nes96$PID)
```

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

Call:  polr(formula = party ~ age + education + income, data = rnes96, Hess = T, method = "logistic")

Coefficients:
         age  education.L  education.Q  education.C  education^4  education^5  education^6
 0.005774902  0.724086814 -0.781360508  0.040168238 -0.019925492 -0.079412657 -0.061103738
      income
 0.012738693

Intercepts:
  Democrat|Independent Independent|Republican
             0.6448794              1.7373541

Residual Deviance: 1984.211
AIC: 2004.211
```

```
summary(pomod)

Call: polr(formula = party ~ age + education + income, data = rnes96, Hess = T, method = "logistic")

Coefficients:
                Value Std. Error  t value
age          0.005775   0.003887  1.48581
education.L  0.724087   0.384388  1.88374
education.Q -0.781361   0.351172 -2.22501
education.C  0.040168   0.291762  0.13767
education^4 -0.019925   0.232429 -0.08573
education^5 -0.079413   0.191533 -0.41462
education^6 -0.061104   0.157747 -0.38735
income       0.012739   0.002140  5.95187

Intercepts:
                       Value   Std. Error t value
Democrat|Independent    0.6449  0.2435     2.6479
Independent|Republican  1.7374  0.2493     6.9694

Residual Deviance: 1984.211
AIC: 2004.211

anova(pomod)

Error in anova.polr(pomod) :
  anova is not implemented for a single "polr" object

fitted(pomod); pomod$fitted; resid(pomod); pomod$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)

Call:    polr(formula = party ~ age + educ.f + income, data = rnes96, Hess = T, method = "logistic")

Coefficients:
           Value Std. Error t value
age     0.005775   0.003887  1.4858
educ.f2 0.582752   0.646411  0.9015
educ.f3 0.998265   0.607858  1.6423
educ.f4 1.222967   0.613844  1.9923
educ.f5 1.152505   0.631241  1.8258
educ.f6 1.166617   0.615485  1.8954
educ.f7 0.836540   0.625520  1.3374
income  0.012739   0.002140  5.9519

Intercepts:
                       Value  Std. Error t value
Democrat|Independent   1.4963 0.6503     2.3008
Independent|Republican 2.5887 0.6537     3.9600

Residual Deviance: 1984.211
AIC: 2004.211

anova(pomod, pomodf)

Likelihood ratio tests of ordinal regression models

Response: party
                     Model Resid. df Resid. Dev   Test    Df     LR stat. Pr(Chi)
1 age + education + income       934   1984.211
2    age + educ.f + income       934   1984.211 1 vs 2     0 9.144742e-09       0
```

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.

```
pomod.aic <- step(pomod, k=2, trace=1)

Start:  AIC=2004.21
party ~ age + education + income

            Df    AIC
- education  6 2002.8
<none>         2004.2
- age        1 2004.4
- income     1 2038.6

Step:  AIC=2002.83
party ~ age + income

         Df    AIC
- age     1 2001.4
<none>      2002.8
- income  1 2047.2

Step:  AIC=2001.36
party ~ income

         Df    AIC
<none>      2001.4
- income  1 2045.3

pomod.bic <- step(pomod, k=log(944.0), trace=1)

Start:  AIC=2052.71
party ~ age + education + income

            Df    AIC
- education  6 2022.2
- age        1 2048.1
<none>         2052.7
- income     1 2082.3

Step:  AIC=2022.23
party ~ age + income

         Df    AIC
- age     1 2015.9
<none>      2022.2
- income  1 2061.8

Step:  AIC=2015.91
party ~ income

         Df    AIC
<none>      2015.9
- income  1 2055.0
##AIC and BIC return the same selection.
```

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

```
anova(pomod, pomod.aic)

Likelihood ratio tests of ordinal regression models

Response: party
```

```
                     Model Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1                   income       941   1995.363
2 age + education + income       934   1984.211 1 vs 2     7 11.15136 0.1321517
```

- It is found that both AIC and BIC select the same model `party` $\sim$ `income` for data `rnes96`.
- The p-value of the likelihood ratio test for comparing `pomod` and `pomod.aic` is 0.132, calculated from the test statistic value of 11.151 and df of 7. Thus, these two models are not significantly different at significance level 0.05 in terms of goodness of fit. But `pomod.aic` is simpler. Therefore, we choose `pomod.aic` as the best model.

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

```
summary(pomod.aic)

Call:
polr(formula = party ~ income, data = rnes96, Hess = T, method = "logistic")

Coefficients:
         Value Std. Error t value
income 0.01312   0.001971   6.657

Intercepts:
                        Value   Std. Error t value
Democrat|Independent     0.2091  0.1123     1.8627
Independent|Republican   1.2916  0.1201    10.7526

Residual Deviance: 1995.363
AIC: 2001.363
```

- The mathematical expression for the best proportional odds model `pomod.aic` is

$$\begin{aligned}\log\frac{\hat{p}_d}{\hat{p}_i+\hat{p}_r} &= \hat{\theta}_1+\hat{\beta}\texttt{income} = 0.2091-0.01312\texttt{income}\\ \log\frac{\hat{p}_d+\hat{p}_i}{\hat{p}_r} &= \hat{\theta}_2+\hat{\beta}\texttt{income} = 1.2916-0.01312\texttt{income}\end{aligned}$$

- We can say that the odds of moving from Democrat to Independent/Republican category (or from Democrat/Independent to Republican) increase by a factor of $\exp(0.01312) = 1.01321$ as income increases by one unit ($1000).

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.

```
predict(pomod.aic, rnes96[944, ],type="probs", se.fit=TRUE)

   Democrat Independent  Republican
  0.2142194   0.2316869   0.5540937
```

```
V<-solve(pomod.aic$Hess)
```

```
                            income Democrat|Independent Independent|Republican
income                  3.883926e-06         0.0001768331           1.476524e-05
Democrat|Independent    1.768331e-04         0.0126018777          -1.018248e-03
Independent|Republican  1.476524e-05        -0.0010182475           3.439770e-03
```

```
rnes96[944, ]
```

```
          party income education age educ.f
944 Independent    115     MAdeg  61      7
```

```
sqrt(V[2,2]+115^2*V[1,1]-2*115*V[1,2]); sqrt(V[3,3]+115^2*V[1,1]-2*115*V[1,3])
```

```
[1] 0.1526276; [1] 0.2267348
```

- For individual 944, `income`=115. So

$$\begin{aligned}\log\frac{\hat{p}_d}{\hat{p}_i+\hat{p}_r} &= 0.2091-0.01312\times 115=-1.299677 \quad \text{with s.e. } 0.1526276\\ \log\frac{\hat{p}_d+\hat{p}_i}{\hat{p}_r} &= 1.2916-0.01312\times 115=-0.217225 \quad \text{with s.e. } 0.2267348\end{aligned}$$

- Thus

$$\begin{aligned}\hat{p}_d &= e^{-1.299677}/(1+e^{-1.299677})=0.2142194\\ \hat{p}_i &= e^{-0.217225}/(1+e^{-0.217225})-\hat{p}_d=0.2316869\\ \hat{p}_r &= 1-\hat{p}_d-\hat{p}_i=0.5540937\end{aligned}$$

- Approx. 95% C.I. for $\theta_1+115\beta$ is $-1.2997\pm 1.96\times 0.1526=(-1.5988,-1.0005)$.
- Approx. 95% C.I. for $\theta_2+115\beta$ is $-0.2172\pm 1.96\times 0.2267=(-0.6616,0.2272)$.
- The approximate 95% C.I. for p_d at `income=115` is

$$\left(\frac{e^{-1.5988}}{1+e^{-1.5988}},\ \frac{e^{-1.0005}}{1+e^{-1.0005}}\right)=(0.1681456,\ 0.2688378)$$

- The approximate 95% C.I. for p_i at `income=115` is

$$\left(\frac{1}{1+e^{-1.0005}}-\frac{1}{1+e^{-0.6616}},\ \frac{1}{1+e^{-1.5988}}-\frac{1}{1+e^{0.2272}}\right)=(0.07153679,\ 0.3884052)$$

- The approximate 95% C.I. for p_r at `income=115` is

$$\left(\frac{1}{1+e^{0.2272}},\ \frac{1}{1+e^{-0.6616}}\right)=(0.4434492,\ 0.6596254)$$