Practice 9 Solutions

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

The `hsb` data was collected from the High School and Beyond Study. We want to see how the relevant variables in the data are related to the choice of `program` with 3 types — academic, vocational, or general — that the students pursue in high school. The response variable `prog` may be regarded as following a multinomial distribution with three levels.

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

```
library(faraway)
help(hsb)
dim(hsb) #n=200, p=11
library(nnet)
head(hsb)
summary(hsb)
summary(hsb$prog)
is.factor(hsb$prog)
help(multinom)
```

2. Fit a trinomial logistic model with `prog` as the response and including 1 as the only predictor (i.e. the null model). Save the results into `hsb0`. Then explore `hsb0` using the commands such as `summary`, `anova`, `fitted`, `prediction`, and `deviance` etc. to see whether you understand the `R` outcomes and are able to interpret them.

```
hsb0 <-multinom(prog~1,data=hsb); hsb0

Call:   multinom(formula = prog ~ 1, data = hsb)
Coefficients:
         (Intercept)
general   -0.8472980
vocation  -0.7419374

Residual Deviance: 408.1933
AIC: 412.1933

summary(hsb0)

Call: multinom(formula = prog ~ 1, data = hsb)

Coefficients:
         (Intercept)
general   -0.8472980
vocation  -0.7419374

Std. Errors:
         (Intercept)
general    0.1781742
vocation   0.1718249
```

```
Residual Deviance: 408.1933
AIC: 412.1933
```

```
anova(hsb0)
```

```
Error in anova.multinom(hsb0) :
  anova is not implemented for a single "multinom" object
```

3. Fit a trinomial logistic model with `prog` as the response and all other variables except `id` as predictors (untransformed, and no interaction terms). Save the results into `hsb1`. Then explore `hsb1` using the commands such as `summary`, `anova`, `fitted`, `prediction`, and `deviance` etc. to see whether you understand the `R` outcomes and are able to interpret them. Also compare `hsb1` with `hsb0` using the `anova` command.

   Note: Change of deviance between two multinomial logit models can still be used to test the difference between the two models, which approximately follows a χ^2 distribution. But the deviance based χ^2 test cannot be used to reliably test the goodness of fit of a multinomial logit model. Other methods are needed.

```
hsb1<-multinom(prog~gender+race+ses+schtyp+read+write+math+science+socst,data=hsb, Hess=T); hsb1
```

```
Call:
multinom(formula = prog ~ gender + race + ses + schtyp + read +
    write + math + science + socst, data = hsb, Hess = T)

Coefficients:
         (Intercept)  gendermale raceasian racehispanic racewhite     seslow
general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156 1.09864111
vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881 0.04747323
         sesmiddle schtyppublic        read       write       math    science
general  0.7029621    0.5845405 -0.04418353 -0.03627381 -0.1092888 0.10193746
vocation 1.1815808    2.0553336 -0.03481202 -0.03166001 -0.1139877 0.05229938
               socst
general  -0.01976995
vocation -0.08040129

Residual Deviance: 305.8705
AIC: 357.8705
```

```
summary(hsb1)
```

```
Call:  multinom(formula = prog ~ gender + race + ses + schtyp + read +
    write + math + science + socst, data = hsb, Hess = T)

Coefficients:
         (Intercept)  gendermale raceasian racehispanic racewhite     seslow
general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156 1.09864111
vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881 0.04747323
         sesmiddle schtyppublic        read       write       math    science
general  0.7029621    0.5845405 -0.04418353 -0.03627381 -0.1092888 0.10193746
vocation 1.1815808    2.0553336 -0.03481202 -0.03166001 -0.1139877 0.05229938
               socst
general  -0.01976995
vocation -0.08040129

Std. Errors:
         (Intercept) gendermale raceasian racehispanic racewhite    seslow
general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763
vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
```

```
         sesmiddle schtyppublic       read      write       math    science
general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441 0.03274038
vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131 0.03424763
              socst
general  0.02712589
vocation 0.02938212

Residual Deviance: 305.8705
AIC: 357.8705
```

```
anova(hsb0,hsb1) #This compares two models based on the chi^2 test.
```

```
Likelihood ratio tests of Multinomial Models

Response: prog
                                                                    Model
1                                                                       1
2 gender + race + ses + schtyp + read + write + math + science + socst
  Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
1       398   408.1933
2       374   305.8705 1 vs 2    24 102.3228 1.203493e-11
```

```
coef(hsb1)
coef(summary(hsb1))
hsb1$fitted
hsb1$residuals
hsb1$deviance
hsb1$edf
hsb1$AIC
hsb1$Hessian
predict(hsb1, hsb[102, ],type="probs")
predict(hsb1, hsb[hsb$id==99, ],type="probs", se.fit=TRUE)
```

4. Perform variable selection based on `hsb1` using `step` function with AIC or BIC option. Save the results into `hsb1.aic` and `hsb1.bic` respectively.

```
hsb1.aic <- step(hsb1, k=2, trace=1)
hsb1.aic
summary(hsb1.aic)

hsb1.bic <- step(hsb1, k=log(200.0), trace=1)
hsb1.bic
```

```
Call:
multinom(formula = prog ~ ses + schtyp + math + science + socst,
    data = hsb, Hess = T)

Coefficients:
         (Intercept)      seslow sesmiddle schtyppublic       math    science
general     2.587029  0.87607389 0.6978995    0.6468812 -0.1212242 0.08209791
vocation    6.687272 -0.01569301 1.2065000    1.9955504 -0.1369641 0.03941237
               socst
general  -0.04441228
vocation -0.09363417

Std. Errors:
         (Intercept)    seslow sesmiddle schtyppublic       math    science
general     1.686492 0.5758781 0.4930330     0.545598 0.03213345 0.02787694
vocation    1.945363 0.6690861 0.5571202     0.812881 0.03591701 0.02864929
              socst
general  0.02344856
vocation 0.02586717

Residual Deviance: 315.5511
AIC: 343.5511
```

It turns out that `hsb1.aic` and `hsb1.bic` are the same.

5. Compare `hsb1` with `hsb1.aic` and `hsb1.bic`.

```
anova(hsb1, hsb1.aic)
summary(hsb1.aic)
```

```
Likelihood ratio tests of Multinomial Models

Response: prog
                                                                    Model
1                                ses + schtyp + math + science + socst
2 gender + race + ses + schtyp + read + write + math + science + socst
  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1       386   315.5511
2       374   305.8705 1 vs 2    12 9.680566 0.6439616
```

Hence, there is no significant difference (p-value = 0.64396) between `hsb1` and `hsb1.aic` (also `hsb1.bic` because `hsb1.aic` and `hsb1.bic` are the same) w.r.t. model goodness of fit. Therefore, we prefer model `hsb1.aic` because it is simplier than `hsb1`.

6. There are two students A and B who have the same math, science and social science scores. Student A comes from a high ses class and private school, while student B comes from a low ses class and public school. Consider the model `hsb1.aic`.

   - Let the probabilities of a student choosing one of the `academic`, `general` and `vocation` programs be p_a, p_g and p_v, respectively.
   - From the R output we see the tri-nomial logit model `hsb1.aic` is estimated to be

$$\begin{aligned}
\log\frac{\hat{p}_g}{\hat{p}_a} &= 2.587 + 0.876\cdot\text{seslow} + 0.698\cdot\text{sesmiddle} + 0.647\cdot\text{schtyppublic} \\
&\quad -0.121\cdot\text{math} + 0.082\cdot\text{science} - 0.044\cdot\text{socst} \\
\log\frac{\hat{p}_v}{\hat{p}_a} &= 6.687 - 0.016\cdot\text{seslow} + 1.207\cdot\text{sesmiddle} + 1.996\cdot\text{schtyppublic} \\
&\quad -0.137\cdot\text{math} + 0.039\cdot\text{science} - 0.094\cdot\text{socst}
\end{aligned}$$

   - The estimated variance-covariance matrix of the MLE of the model's regression coefficients can be obtained from R command `V <- solve(hsb1.aic$Hessian)`, which is a 14×14 matrix.

   (a) Estimate the odds ratio of choosing `general` program against `academic` program for student A versus student B. Find an approximate 95% confidence interval for this odds ratio.

   -

$$\log\left[\frac{\hat{p}_g}{\hat{p}_a}\right]_A - \log\left[\frac{\hat{p}_g}{\hat{p}_a}\right]_B = -0.876 - 0.647 = -1.523 \quad \text{with s.e. } 0.745.$$

```
sqrt(c(-1,-1)%*%V[c(2,4),c(2,4)]%*%c(-1,-1))
```

```
0.7451041
```

- The referenced odds ratio is estimated to be $e^{-1.523} = 0.218$, with the approximate 95% C.I.

$$e^{-1.523 \pm 1.96 \cdot 0.745} = (e^{-2.9832},\ e^{-0.0628}) = (0.0506, 0.9391).$$

(b) Estimate the odds ratio of choosing `vocation` program against `academic` program for student A versus student B. Find an approximate 95% confidence interval for this odds ratio.

- $$\log\left[\frac{\hat{p}_v}{\hat{p}_a}\right]_A - \log\left[\frac{\hat{p}_v}{\hat{p}_a}\right]_B = 0.016 - 1.996 = -1.980 \quad \text{with s.e. } 1.001.$$

```
sqrt(c(-1,-1)%*%V[c(9,11),c(9,11)]%*%c(-1,-1))
1.001
```

- The referenced odds ratio estimate is $e^{-1.980} = 0.138$, with the approximate 95% C.I.

$$e^{-1.980 \pm 1.96 \cdot 1.001} = (e^{-3.9426},\ e^{-0.0174}) = (0.0194, 0.9827).$$

(c) Estimate the odds ratio of choosing `general` program against `vocation` program for student A versus student B. Find an approximate 95% confidence interval for this odds ratio.

- $$\log\left[\frac{\hat{p}_g}{\hat{p}_v}\right]_A - \log\left[\frac{\hat{p}_g}{\hat{p}_v}\right]_B = -(0.876+0.016)-(0.647-1.996) = 0.457 \quad \text{with s.e. } 1.053.$$

```
sqrt(c(-1,-1,1,1)%*%V[c(2,4,9,11),c(2,4,9,11)]%*%c(-1,-1,1,1))
1.053158
```

- The referenced odds ratio is estimated to be $e^{0.457} = 1.579$, with the approximate 95% C.I.

$$e^{0.457 \pm 1.96 \cdot 1.053} = (e^{-1.6072},\ e^{2.5212}) = (0.2005, 12.4434).$$

7. For the student with id 99, compute the predicted probabilities of the three possible choices based on the best model among `hsb1`, `hsb1.aic` and `hsb1.bic`.

```
hsb[hsb$id==99,]
```

```
    id gender  race  ses schtyp    prog read write math science socst
102 99 female white high public general   47    59   56      66    61
```

```
predict(hsb1.aic, hsb[hsb$id==99, ],type="probs", se.fit=TRUE)
```

```
  academic    general   vocation
0.64426309 0.27665609 0.07908082
```