7.9 (Odds ratios for coronary disease data):

(a) The odds ratio is $\exp(0.035) = 1.0356$. The 95% CI on β for Sex is $0.035 \pm 1.96 \times 0.0148 = [0.006, 0.064]$ Hence the 95% CI on the odds ratio $\exp(\beta)$ is $[\exp(0.006), \exp(0.064)] = [1.006, 1.066]$.

(b) The odds ratio equals
$\exp(0.0906 \times 10)\exp(0.0755 \times 20)\exp(0.035) = \exp(2.451) = 11.60$.
Therefore the odds of coronary disease for this male are

$$\frac{1}{10} \times 11.60 = 1.160.$$

If p is the probability of coronary disease for this male then

$$\frac{p}{1-p} = 1.160 \quad \Longrightarrow \quad p = \frac{1.160}{1+1.160} = 0.5370.$$

7.12 (Pregnancy duration):

(a) We will do this exercise using both `nnet` and `mlogit`. Both give the same results for this data.

```
library(nnet)
preg = read.csv("c:/data/Pregnancy.csv")
preg$Age1[1:nrow(preg)] = 0
preg$Age1[which(preg$Age==1)] = 1
preg$Age3[1:nrow(preg)] = 0
preg$Age3[which(preg$Age==3)] = 1
evenrow = seq(2,nrow(preg),2)
oddrow = seq(1,nrow(preg),2)
train = preg[oddrow,]
test = preg[evenrow,]
lgfit1 = multinom(Duration~Nutrition+Alcohol+Smoking+Age1+Age3,
+       data=train)
summary(lgfit1)
trainpred=predict(lgfit1, newdata=train);
table(train$Duration, trainpred)
testpred= predict(lgfit1, newdata=test)
```

```
table(test$Duration, testpred)
```

The output using "Preterm" as the reference category is shown below.

```
Coefficients:
  (Intercept)  Nutrition   Alcohol    Smoking      Age1       Age3
2   -1.085473 0.01553680 -1.060328 -0.6273106  1.006311  0.4746559
3   -1.862191 0.03713018 -2.113399 -2.7174780 -2.361983 -0.7404409

Std. Errors:
  (Intercept)  Nutrition   Alcohol   Smoking     Age1     Age3
2    2.521758 0.01876176 0.8216689 0.8575421 1.025365 1.075981
3    2.657861 0.02022154 0.9752696 0.9989745 1.504805 1.329033

Residual Deviance: 85.06886
AIC: 109.0689
```

The following output using the `mlogit` library agrees with the above output but also gives large sample z-statistics.

```
Coefficients :
               Estimate Std. Error t-value Pr(>|t|)
2:(intercept) -1.085419   2.521756 -0.4304 0.666889
3:(intercept) -1.862345   2.657872 -0.7007 0.483496
2:Nutrition    0.015536   0.018762  0.8281 0.407625
3:Nutrition    0.037131   0.020222  1.8362 0.066328 .
2:Alcohol     -1.060329   0.821668 -1.2905 0.196891
3:Alcohol     -2.113373   0.975271 -2.1670 0.030238 *
2:Smoking     -0.627292   0.857542 -0.7315 0.464474
3:Smoking     -2.717454   0.998979 -2.7202 0.006524 **
2:Age1         1.006349   1.025369  0.9815 0.326370
3:Age1        -2.361864   1.504796 -1.5696 0.116518
2:Age3         0.474632   1.075975  0.4411 0.659128
3:Age3        -0.740482   1.329040 -0.5572 0.577421
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -42.534
McFadden R^2:  0.22997
Likelihood ratio test : chisq = 25.406 (p.value = 0.0046268)
```

We see that only the Alcohol and Smoking coefficients for log-odds of full term vs. preterm pregnancy are significant and both have negative signs. This means that if the mother is an alcohol user or a smoker then

the odds of a full term pregnancy w.r.t. the odds of a preterm pregnancy decrease or mother is less likely to have a full term pregnancy.
The classification table is as follows.

```
  testpred
     1  2  3
  1  4  6  3
  2  4  9  4
  3  2  3 16
```

So the CCR is $(4+9+16)/51 = 29/51 = 56.86\%$.

(b) Ordinal logistic regression model is fitted using the following R code.

```
library(MASS)
preg = read.csv("c:/data/Pregnancy.csv")
train = preg[oddrow,]
train$Duration.ordered = as.ordered(train$Duration)
test = preg[evenrow,]
test$Duration.ordered = as.ordered(test$Duration)
lgfit2 = polr(Duration.ordered~Nutrition+Alcohol+Smoking+Age1+Age3,
+         data=train)
summary(lgfit2)
test.prob = predict(lgfit2, newdata=test)
table(test$Duration, test.prob)
```

The output is shown below.

```
Coefficients:
             Value Std. Error t value
Nutrition  0.02954    0.01384  2.1350
Alcohol   -1.62122    0.64720 -2.5050
Smoking   -1.80054    0.61526 -2.9264
Age1      -1.37253    0.74461 -1.8433
Age3      -0.54337    0.79352 -0.6848

Intercepts:
    Value   Std. Error t value
1|2  0.5350  1.8214     0.2937
2|3  2.5610  1.8523     1.3826

Residual Deviance: 91.30409
AIC: 105.3041
```

We see that except for Age3, other predictors are significant or nearly significant. Alcohol, Smoking and Age1 have negative signs. But since

R fits the model

$$P(y \leq k) = \frac{\exp(\beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\beta_{0k} - \boldsymbol{x}'\boldsymbol{\beta})},$$

the signs are reversed. Thus if a mother is an alcohol user or a smoker then the probability that the pregnancy outcome is 1 (preterm) or 2 (intermediate term) increases. On the other hand better nutrition would reduce this probability.
The classification table is as follows.

```
test.prob
   1  2  3
1  5  5  3
2  3 10  4
3  2  0 19
```

So the CCR is $(5 + 10 + 19)/51 = 34/51 = 66.67\%$. So ordinal logistic regression gives a significant improvement over nominal logistic regression.

7.13 (Mammography testing history):

(a) Nominal logistic regression model is fitted using the following R code. Note that I recoded ME=0,1,2 to ME=0,2,1 to represent natural ordering for ordinal logistic regression. The reference category is "Never".

```
library(nnet)
mamgraph = read.csv("c:/data/Mammography.csv")
evenrow = seq(2,nrow(mamgraph),2)
oddrow = seq(1,nrow(mamgraph),2)
train = mamgraph[oddrow,]
test = mamgraph[evenrow,]
fit1 = multinom(ME~PB+HIST, data=train)
summary(fit1)
testpred= predict(fit1, newdata=test)
table(test$ME, testpred)
```

The output is as follows.

```
multinom(formula = ME ~ PB + HIST, data = train)

Coefficients:
  (Intercept)          PB     HIST
1   0.2424622 -0.2250368 1.064849
2   1.0356672 -0.2996926 1.655385
```

```
Std. Errors:
  (Intercept)         PB       HIST
1   0.7918821 0.10642547 0.6300210
2   0.7153200 0.09834596 0.5267861

Residual Deviance: 363.2462
AIC: 375.2462
  testpred
      0   1   2
  0 106   0   4
  1  37   0   4
  2  48   0   7
```

The CCR using this model is $(106 + 7)/206 = 54.85\%$.
The following output using the `mlogit` library agrees with the above output but gives also large sample z-statistics.

```
Coefficients :
               Estimate Std. Error t-value Pr(>|t|)
1:(intercept)  0.242514   0.791882  0.3062 0.759414
2:(intercept)  1.035712   0.715321  1.4479 0.147646
1:PB          -0.225043   0.106426 -2.1146 0.034468 *
2:PB          -0.299698   0.098346 -3.0474 0.002308 **
1:HIST         1.064827   0.630023  1.6901 0.091001 .
2:HIST         1.655380   0.526786  3.1424 0.001676 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -181.62
McFadden R^2:  0.06256
Likelihood ratio test : chisq = 24.241 (p.value = 7.1462e-05)
```

The two fitted models are:

$$\ln\left[\frac{p_1}{p_0}\right] = 0.2425 - 0.2250\text{PB} + 1.0648\text{HIST}$$

and

$$\ln\left[\frac{p_2}{p_0}\right] = 1.0357 - 0.2997\text{PB} + 1.6554\text{HIST}.$$

The signs of the regression coefficients in these models accord with our intuition. Since low values of PB mean higher perceived benefit of mammography, negative coefficients mean that lower PB values result in higher probabilities of outcome 1 (Duration $>$ 1 year) and outcome

2 (Duration $\leq$ 1 year). Similarly if there is family history of breast cancer (HIST = 1) then these probabilities are higher because the coefficient of HIST is positive in both models.

(b) Ordinal logistic regression model is fitted using the following R code.

```
library(MASS)
mamgraph = read.csv("c:/data/Mammography.csv")
mamgraph$ME.ordered = as.ordered(preg$ME)
train = mamgraph[oddrow,]
train$ME.ordered = as.ordered(train$ME)
test = mamgraph[evenrow,]
test$ME.ordered = as.ordered(test$ME)
fit2 = polr(ME.ordered~PB+HIST, data=train)
summary(fit2)
test.prob = predict(fit2, newdata=test)
table(test$ME, test.prob)

The fitted ordinal regression model is
\begin{verbatim}
Coefficients:
       Value Std. Error t value
PB   -0.2626    0.07728  -3.398
HIST  1.3724    0.42023   3.266

Intercepts:
    Value   Std. Error t value
0|1 -1.3439  0.5746    -2.3389
1|2 -0.5102  0.5699    -0.8952

Residual Deviance: 363.2292
AIC: 371.2292
```

The classification table is exactly the same as that obtained for nominal logistic regression. So there is no gain in correct classification by using ordinal logistic regression.

```
test.prob
      0   1   2
  0 106   0   4
  1  37   0   4
  2  48   0   7
```

7.14 (Program choices by high school students)

(a) We first fitted a full model using `mlogit` and the reference outcome being academic, resulting in the following output. You may use `nnet` but then you will need to calculate the z-statistics and their P-values by hand. To avoid calculating the P-values, you may fix $\alpha = .05$ and the corresponding two-sided critical value of z as 1.96. Then any $|z| < 1.96$ is not significant at $\alpha = .05$.

```
Coefficients :
                        Estimate Std. Error z-value Pr(>|z|)
general:(intercept)     2.741182   1.689608  1.6224 0.104723
vocation:(intercept)    6.352443   1.944359  3.2671 0.001086 **
general:gendermale     -0.132739   0.450633 -0.2946 0.768329
vocation:gendermale    -0.342915   0.484328 -0.7080 0.478931
general:seslow          1.032056   0.569374  1.8126 0.069891 .
vocation:seslow         0.344287   0.655035  0.5256 0.599166
general:sesmiddle       0.684560   0.497530  1.3759 0.168847
vocation:sesmiddle      1.158923   0.551030  2.1032 0.035449 *
general:schtyppublic    0.550113   0.550423  0.9994 0.317583
vocation:schtyppublic   1.890278   0.810721  2.3316 0.019722 *
general:read           -0.052245   0.029109 -1.7948 0.072685 .
vocation:read          -0.056302   0.032344 -1.7407 0.081734 .
general:write          -0.034122   0.031456 -1.0847 0.278034
vocation:write         -0.061915   0.032699 -1.8935 0.058292 .
general:math           -0.098836   0.033799 -2.9242 0.003454 **
vocation:math          -0.122845   0.037336 -3.2903 0.001001 **
general:science         0.100159   0.030676  3.2651 0.001094 **
vocation:science        0.060421   0.031555  1.9148 0.055515 .
```

We see that gender, read and write are not statistically significant. We refitted the model by dropping these nonsignificant predictors resulting in the following output.

```
Coefficients :
                        Estimate Std. Error z-value  Pr(>|z|)
general:(intercept)     1.336282   1.516229  0.8813  0.378145
vocation:(intercept)    4.423654   1.739430  2.5432  0.010985 *
general:seslow          1.076890   0.556405  1.9354  0.052936 .
vocation:seslow         0.363031   0.633669  0.5729  0.566711
general:sesmiddle       0.739685   0.488611  1.5139  0.130063
vocation:sesmiddle      1.221593   0.535179  2.2826  0.022455 *
general:schtyppublic    0.594393   0.540696  1.0993  0.271632
vocation:schtyppublic   2.021408   0.806696  2.5058  0.012218 *
```

```
general:math          -0.133029   0.031256 -4.2561 2.080e-05 ***
vocation:math         -0.166843   0.034508 -4.8349 1.332e-06 ***
general:science        0.071992   0.026679  2.6984  0.006967 **
vocation:science       0.017480   0.026679  0.6552  0.512338
```

Now all variables are significant for at least one categorical outcome. The interpretation of the coefficients is as follows. The coefficients of seslow are both positive implying that low ses students are more likely to choose general or vocational program. The same is true for middle ses students. Thus only high ses students tend to select academic program. Similarly public school students opt for general or vocational program, so private chool students opt for academic program. As expected, high math scores encourage students to choose academic programs, but surprisingly high science scores have the opposite effect.

(b) The predicted probabilities for the three program choices are calculated as follows.

```
> testdata0 = data.frame('ses' = 'high', 'schtyp' = 'private',
'math' = 52, > 'science' = 53)
> predict(fit0,type='probs',newdata=testdata0)
 academic   general  vocation
0.8284860 0.1417333 0.0297807
```

The most likely program choice is academic.

(c) The ordinal regression model was fitted with program choice order specified as vocation, general and academic as can be seen below in the R output. Only the gender variable was nonsignificant. The fitted model omitting gender variable is as follows.

```
> choice$prog2= ordered(choice$prog,levels=c('vocation',
'general','academic'),
> labels=c('vocation','general','academic'))
> fit2=clm(prog2~ses+schtyp+read+write+math+science,data=choice)
> summary(fit2)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
seslow       -0.42054    0.44497  -0.945 0.344609
sesmiddle    -0.89023    0.38802  -2.294 0.021774 *
schtyppublic -1.12568    0.47497  -2.370 0.017787 *
read          0.04873    0.02238   2.177 0.029493 *
write         0.04064    0.02093   1.942 0.052155 .
math          0.09883    0.02551   3.874 0.000107 ***
science      -0.05306    0.02193  -2.420 0.015539 *
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Threshold coefficients:
                 Estimate Std. Error z value
vocation|general    4.102      1.278   3.211
general|academic    5.490      1.308   4.197
```

Note that the signs of the coefficients are reversed in the above fitted model. Thus being in low or middle ses group increases the chances of choosing vocation or general program; similarly for public school. On the other hand, having higher read, write or math scores increases the chances of choosing academic program. Surprisingly, science scores have the opposite effect, i.e., a higher science score makes it more likely to choose vocation or general program.

The predicted probabilities for the three program choices are calculated as follows.

```
> testdata2 = data.frame('ses' = 'high', 'schtyp' =
'private','read'=50, > 'write'=54, 'math' = 52,
'science' = 53)
> predict(fit2,type='prob',newdata=testdata2)
$`fit`
    vocation   general  academic
1 0.0543978 0.1327919 0.8128103
```

The predicted probabilities are similar to those obtained with the the nominal logistic regression model. The most likely program choice is academic.

(d) The confusion matrix using the nominal regression model is as follows.

```
          Y.hat.1
             1  2  3
  academic 86  7 12
  general  24 11 10
  vocation 15  6 29
> correct.rate0
[1] 0.63
```

Thus the CCR is 63%.

The confusion matrix using the ordinal regression model is as follows.

```
           Y.hat.2
            academic general vocation
  academic        94       0       11
```

```
  general         26        0        19
  vocation        17        0        33
> correct.rate2 = sum(diag(ctable2)[1:3])/n;
> correct.rate2
[1] 0.635
```

Although the CCR (63.5%) is nearly the same as for the nominal regression model, note that no students are classified to the general program choice.