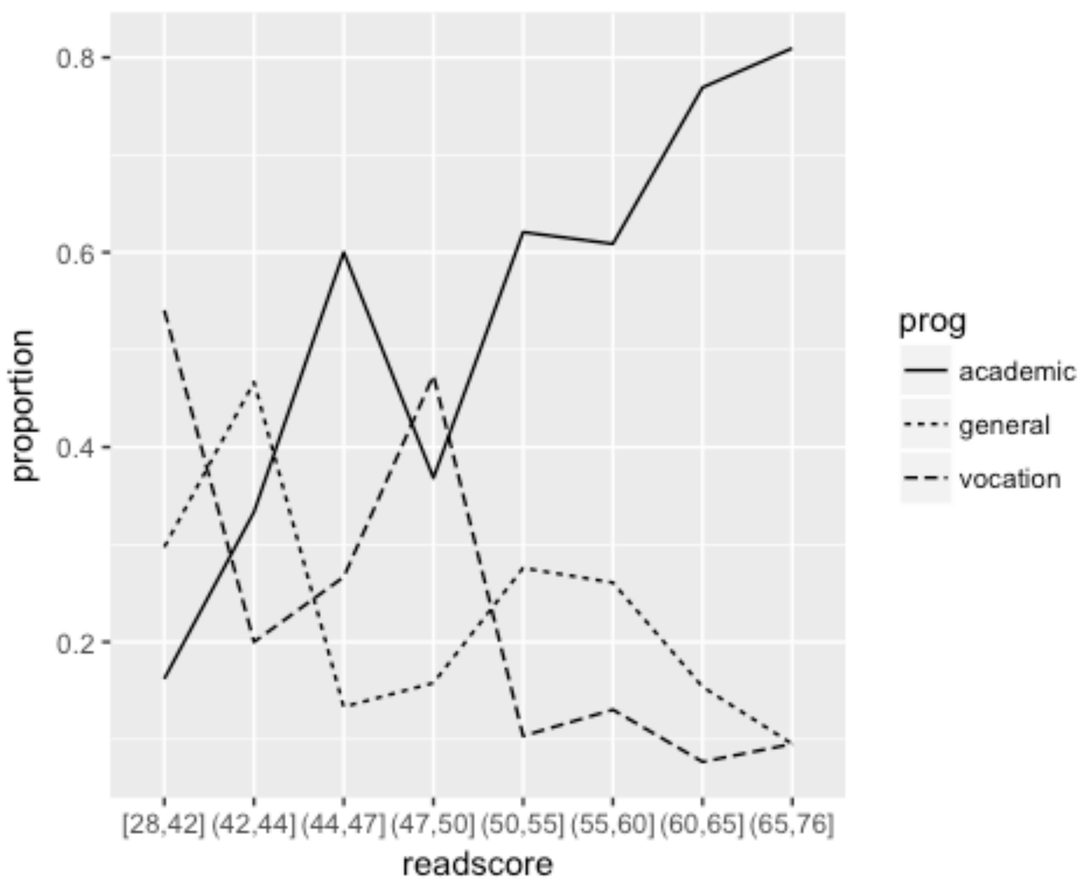
NAEEM-NOWROUZI-707-HW3

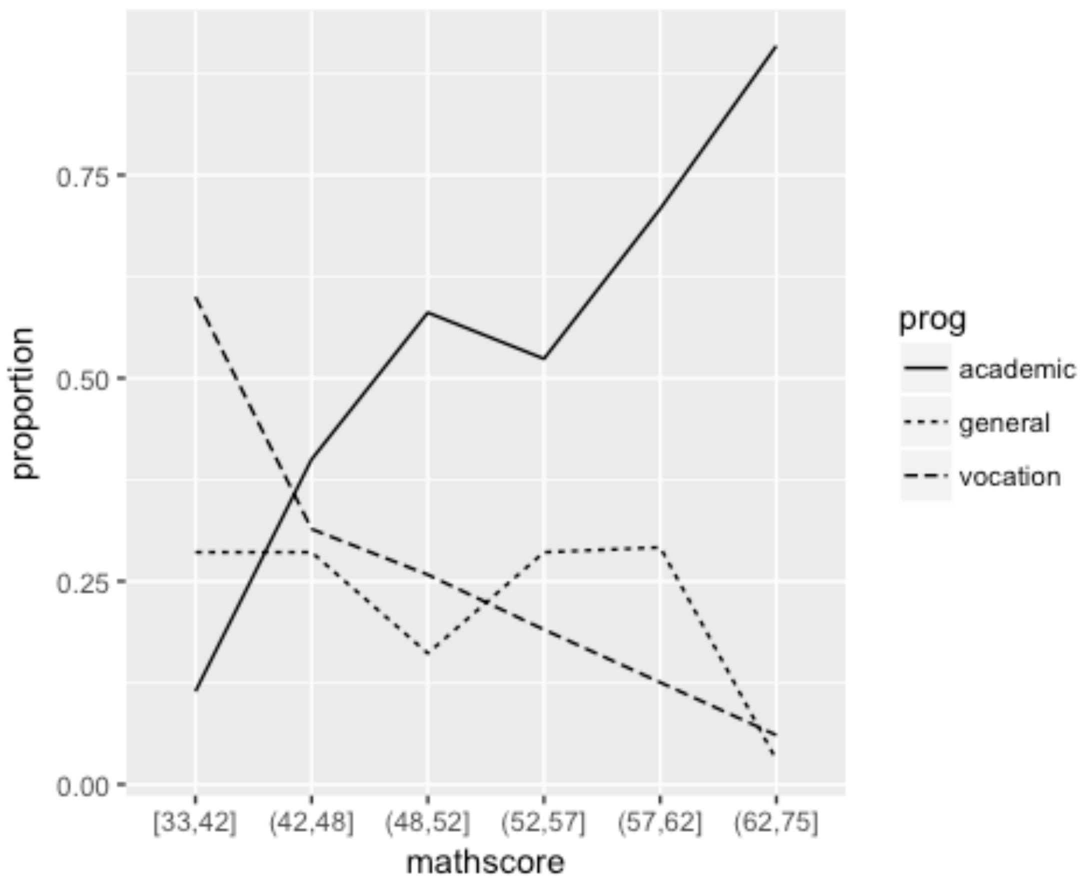
Chapter 7, Problem 1

```
library(faraway)
data(hsb)
hsbm <- na.omit(hsb)

# 1.a) Produce a table showing the proportion of males and females choosing
# the three different programs.
library(tidyverse)

prop.gen <- group_by(hsb, gender, prog) %>% summarise(count = n()) %>%
           group_by(gender) %>% mutate(gtotal = sum(count), proportion =
count/gtotal)
prop.gen

## # A tibble: 6 x 5
## # Groups:   gender [2]
##   gender prog     count gtotal proportion
##   <fct>  <fct>    <int>  <int>      <dbl>
## 1 female academic    58    109      0.532
## 2 female general     24    109      0.220
## 3 female vocation    27    109      0.248
## 4 male   academic    47     91      0.516
## 5 male   general     21     91      0.231
## 6 male   vocation    23     91      0.253

# We see that within the females, 53% choose academic program type, 22%
choose general, and 25% choose vocational. And withing the males, 52% choose
academic, 23% choose general, and 25% choose vocational.

# Do the same for SES
prop.ses <- group_by(hsb, ses, prog) %>% summarise(count = n()) %>%
           group_by(ses) %>% mutate(gtotal = sum(count), proportion =
count/gtotal)
prop.ses

## # A tibble: 9 x 5
## # Groups:   ses [3]
##   ses    prog     count gtotal proportion
##   <fct>  <fct>    <int>  <int>      <dbl>
## 1 high   academic    42     58      0.724
## 2 high   general      9     58      0.155
## 3 high   vocation     7     58      0.121
## 4 low    academic    19     47      0.404
## 5 low    general     16     47      0.340
## 6 low    vocation    12     47      0.255
```

```
## 7 middle academic    44     95      0.463
## 8 middle general     20     95      0.211
## 9 middle vocation    31     95      0.326

# We observe that in the high socioeconomic class 72% choose the academic
program, 16% choose the general program, and 12% choose the vocational
program. In the middle class, 46% choose academic, 21% choose general, and
33% choose vocational. In the low class, 40% choose academic, 34% choose
general, and 26% choose vocational.


# 1.b) Plot the relationship between program choice and reading score.

progread <- mutate(hsbm, readscore = cut_number(read, 8)) %>%
            group_by(readscore, prog) %>% summarise(count = n()) %>%
            group_by(readscore) %>% mutate(tot = sum(count), proportion =
count/tot)

ggplot(progread, aes(x = readscore, y = proportion, group = prog, linetype =
prog)) +
  geom_line()
```



```r
# Repeat the above plot for math scores

progmath <- mutate(hsbm, mathscore = cut_number(math, 6)) %>%
       group_by(mathscore, prog) %>% summarise(count = n()) %>%
       group_by(mathscore) %>% mutate(tot = sum(count), proportion =
count/tot)

ggplot(progmath, aes(x = mathscore, y = proportion, group = prog, linetype =
prog)) +
  geom_line()
```




```r
# We can see in the plots that for both meath and reading scores, as they
increase the proportion of academic program choice increases and the other
two program type decrease.

# 1.c) Compute the correlation matrix for the five subject scores.

cor(hsb[7:11])

##              read     write      math   science     socst
## read    1.0000000 0.5967765 0.6622801 0.6301579 0.6214843
## write   0.5967765 1.0000000 0.6174493 0.5704416 0.6047932
## math    0.6622801 0.6174493 1.0000000 0.6307332 0.5444803
```

```
## science 0.6301579 0.5704416 0.6307332 1.0000000 0.4651060
## socst   0.6214843 0.6047932 0.5444803 0.4651060 1.0000000

# 1.d)

library(nnet)

multnommod0 <- multinom(prog ~ ., hsbm)

## # weights:  45 (28 variable)
## initial  value 219.722458
## iter  10 value 181.098338
## iter  20 value 154.577078
## iter  30 value 152.478856
## final  value 152.478368
## converged

summary(multnommod0)

## Call:
## multinom(formula = prog ~ ., data = hsbm)
##
## Coefficients:
##          (Intercept)           id  gendermale  raceasian racehispanic
## general     4.263658 -0.007332836 -0.04666403  1.2170225   -0.8702109
## vocation    7.845921 -0.003680462 -0.29724832 -0.7863428   -0.3236628
##          racewhite    seslow sesmiddle schtyppublic        read
## general  0.8609754 1.1547399 0.7430976    0.1384853 -0.05445264
## vocation 0.6223190 0.0728241 1.1897765    1.8285649 -0.04078359
##                write       math   science       socst
## general  -0.03716360 -0.1037470 0.1065258 -0.01786542
## vocation -0.03220268 -0.1099712 0.0537472 -0.07959798
##
## Std. Errors:
##          (Intercept)          id gendermale raceasian racehispanic
## general     1.960941 0.007678009  0.4587870  1.064969    0.9286986
## vocation    2.288984 0.008408855  0.5048241  1.476435    0.8924359
##          racewhite    seslow sesmiddle schtyppublic       read      write
## general  0.9438010 0.6134530 0.5096129    0.7338284 0.03300204 0.03398842
## vocation 0.9519097 0.7067682 0.5739217    0.9981540 0.03583547 0.03597627
##                math    science      socst
## general  0.03556357 0.03331314 0.02737227
## vocation 0.03885464 0.03445137 0.02963317
##
## Residual Deviance: 304.9567
## AIC: 360.9567

# The variable math has unusual coefficients. We can see from the correlation
matrix and the above plots that math score has the largest effect on the
outcome.
```

```
# 1.e)

hsbm1 <- mutate(hsb, scoresum = read + write + math + science + socst)
hsbm1 <- hsbm1[-c(7:11)]
multnommod1 <- multinom(prog ~ ., hsbm1)

## # weights:  33 (20 variable)
## initial  value 219.722458
## iter  10 value 167.158173
## iter  20 value 164.141699
## final  value 164.130704
## converged

summary(multnommod1)

## Call:
## multinom(formula = prog ~ ., data = hsbm1)
##
## Coefficients:
##          (Intercept)           id  gendermale  raceasian racehispanic
## general     3.227335 -0.003708235  0.24883040  1.0243408   -0.5484976
## vocation    7.112010 -0.003220142 -0.09614882 -0.6015843   -0.1937564
##          racewhite    seslow sesmiddle schtyppublic    scoresum
## general   1.060033 1.0593830 0.6350558    0.3875245 -0.02052599
## vocation  1.098265 0.2517821 1.1874930    1.8098161 -0.04125543
##
## Std. Errors:
##          (Intercept)          id gendermale raceasian racehispanic
## general     1.798815 0.006823237  0.3941480 0.9439661    0.8799224
## vocation    2.157426 0.007659938  0.4364287 1.3769618    0.8411264
##          racewhite    seslow sesmiddle schtyppublic    scoresum
## general  0.8740777 0.5664146 0.4789630    0.6826598 0.005976099
## vocation 0.8970833 0.6797684 0.5566371    0.9568939 0.007225491
##
## Residual Deviance: 328.2614
## AIC: 368.2614

# We can see that the first model that includes the scores seperately has a
lower residual deviance and lower AIC, so we conclude that it fits better.


# 1.f)

# Use a stepwise method to reduce the model

bestmodel <- step(multnommod0, trace = 0)

## trying - id
## trying - gender
```

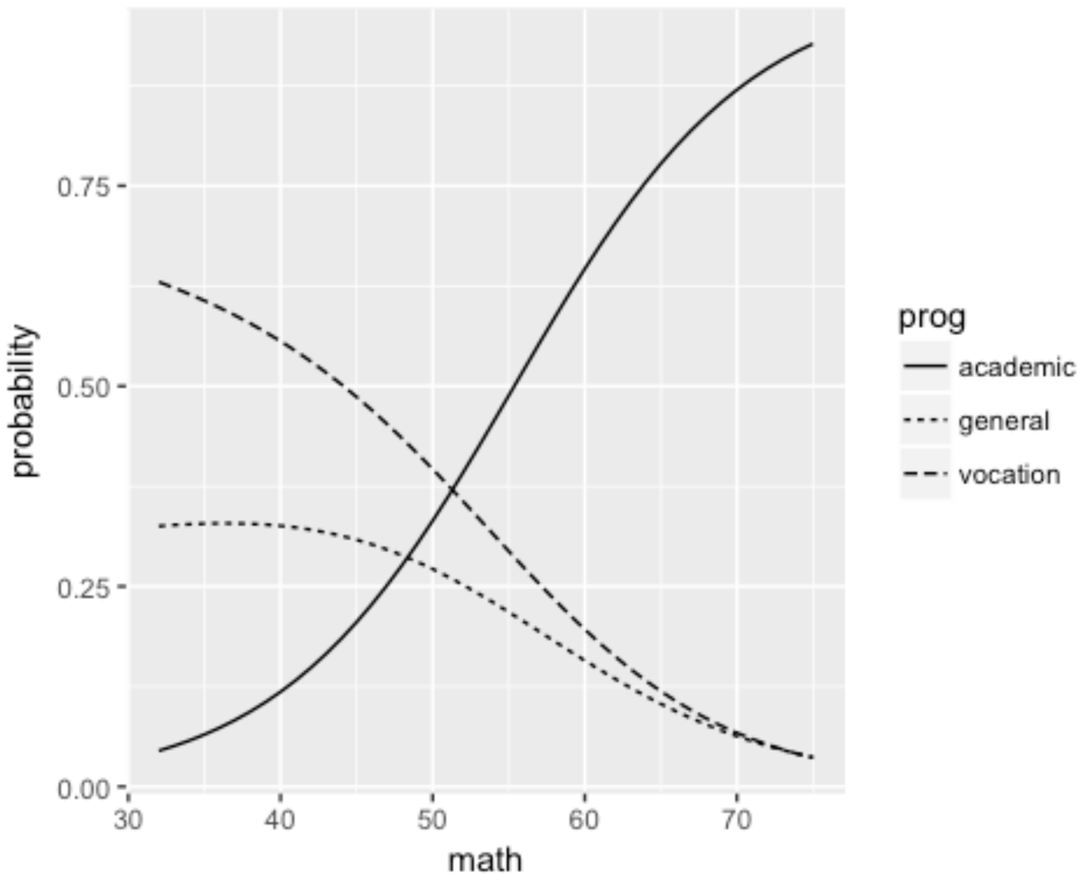
```
## trying - race
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
## # weights:  36 (22 variable)
## initial  value 219.722458
## iter  10 value 180.513235
## iter  20 value 156.699696
## final  value 155.608810
## converged
## trying - id
## trying - gender
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
## # weights:  33 (20 variable)
## initial  value 219.722458
## iter  10 value 172.925326
## iter  20 value 156.065379
## final  value 155.776076
## converged
## trying - gender
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
## # weights:  30 (18 variable)
## initial  value 219.722458
## iter  10 value 172.662548
## iter  20 value 156.063823
## final  value 156.032828
## converged
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
```

```
## # weights:  27 (16 variable)
## initial  value 219.722458
## iter  10 value 176.827677
## iter  20 value 156.410686
## final  value 156.406678
## converged
## trying - ses
## trying - schtyp
## trying - read
## trying - math
## trying - science
## trying - socst
## # weights:  24 (14 variable)
## initial  value 219.722458
## iter  10 value 171.169761
## iter  20 value 157.775586
## final  value 157.775540
## converged
## trying - ses
## trying - schtyp
## trying - math
## trying - science
## trying - socst
```

```
summary(bestmodel)
```

```
## Call:
## multinom(formula = prog ~ ses + schtyp + math + science + socst,
##     data = hsbm)
##
## Coefficients:
##          (Intercept)      seslow sesmiddle schtyppublic       math
## general     2.587029  0.87607389 0.6978995    0.6468812 -0.1212242
## vocation    6.687272 -0.01569301 1.2065000    1.9955504 -0.1369641
##             science       socst
## general  0.08209791 -0.04441228
## vocation 0.03941237 -0.09363417
##
## Std. Errors:
##          (Intercept)    seslow sesmiddle schtyppublic       math
## general     1.686492 0.5758781 0.4930330     0.545598 0.03213345
## vocation    1.945363 0.6690861 0.5571202     0.812881 0.03591701
##             science      socst
## general  0.02787694 0.02344856
## vocation 0.02864929 0.02586717
##
## Residual Deviance: 315.5511
## AIC: 343.5511
```

```r
# Since this method is based on AIC we see that the reduced model with five
variable (ses + schtyp + math + science + socst) has a lower AIC, however,
the deviance is slightly higher, we proceed with this reduced model.



# 1.g)

# Define the observed range of math scores
mathscorlev <- 32:75

# Find the most common levels of the factors in the model and the mean of the
# other predictors in the model.
summary(hsbm$ses)

##   high    low middle
##     58     47     95

summary(hsbm$schtyp)

## private  public
##      32     168

# Get the pridected values for the observed range of math scores.
preds <- data.frame(math=mathscorlev,
                    predict(bestmodel, newdata = data.frame(ses = "middle",
                          schtyp = "public", math = mathscorlev, science =
mean(hsbm$science),
                           socst = mean(hsbm$socst)),type = "probs"))

library(tidyr)

lpred <- gather(preds, prog, probability, -math)

ggplot(lpred, aes(x = math, y = probability, group = prog, linetype = prog))
+ geom_line()
```




```
#lpred

# Clearly, the probability of choosing the academic program type increases
rapidly as the math scores get higher, and the other two programs type have a
decreased probability  of begin chosen as the math score goes higher.


# 1.h)

# Compute a table
# data.frame(ses = hsbm$ses, schtyp = hsbm$schtyp,
 #          predict(bestmodel,newdata = data.frame(ses = hsbm$ses,
  #         schtyp = hsbm$schtyp, math = mean(hsbm$math), science =
mean(hsbm$science),
   #        socst = mean(hsbm$socst)), type = "probs"))

xtabs(predict(bestmodel,newdata = data.frame(ses = hsbm$ses,
           schtyp = hsbm$schtyp, math = mean(hsbm$math), science =
mean(hsbm$science),
           socst = mean(hsbm$socst)), type = "probs") ~ hsbm$schtyp,
hsbm$ses)
```

```
##
## hsbm$schtyp  academic   general  vocation
##     private 24.018832  6.112748  1.868419
##     public  84.782164 43.954711 39.263126
```

```
# 1.i)

# The student with id 99 is at row 102 in the data set.
predict(bestmodel, newdata = hsb[102,])
```

```
## [1] academic
## Levels: academic general vocation
```

```
# The predicted value is academic which is wrong, the correct program type is
general.

# 1.j) Construct a table of predicted and observed values.

xtabs(~ predict(bestmodel) + hsbm$prog)
```

```
##                   hsbm$prog
## predict(bestmodel) academic general vocation
##           academic       87      22       17
##           general         7      10        4
##           vocation       11      13       29
```

```
# Compute the correct classification rate

(87 + 10 + 29) / nrow(hsbm)
```

```
## [1] 0.63
```

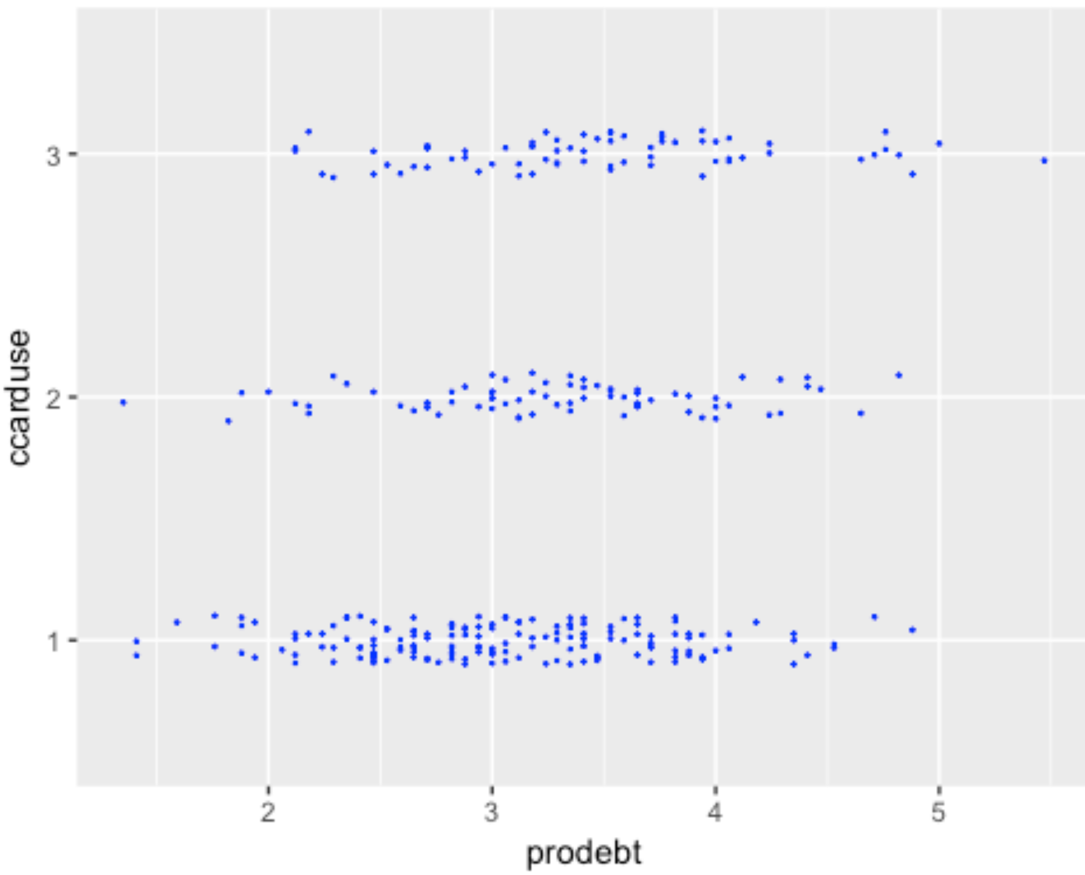
```
# We see that 63% of the data are correctly classified, which is not
impressive given that we expect the model to perform worse than this on new
observations.
```

Problem 5, Chapter 7

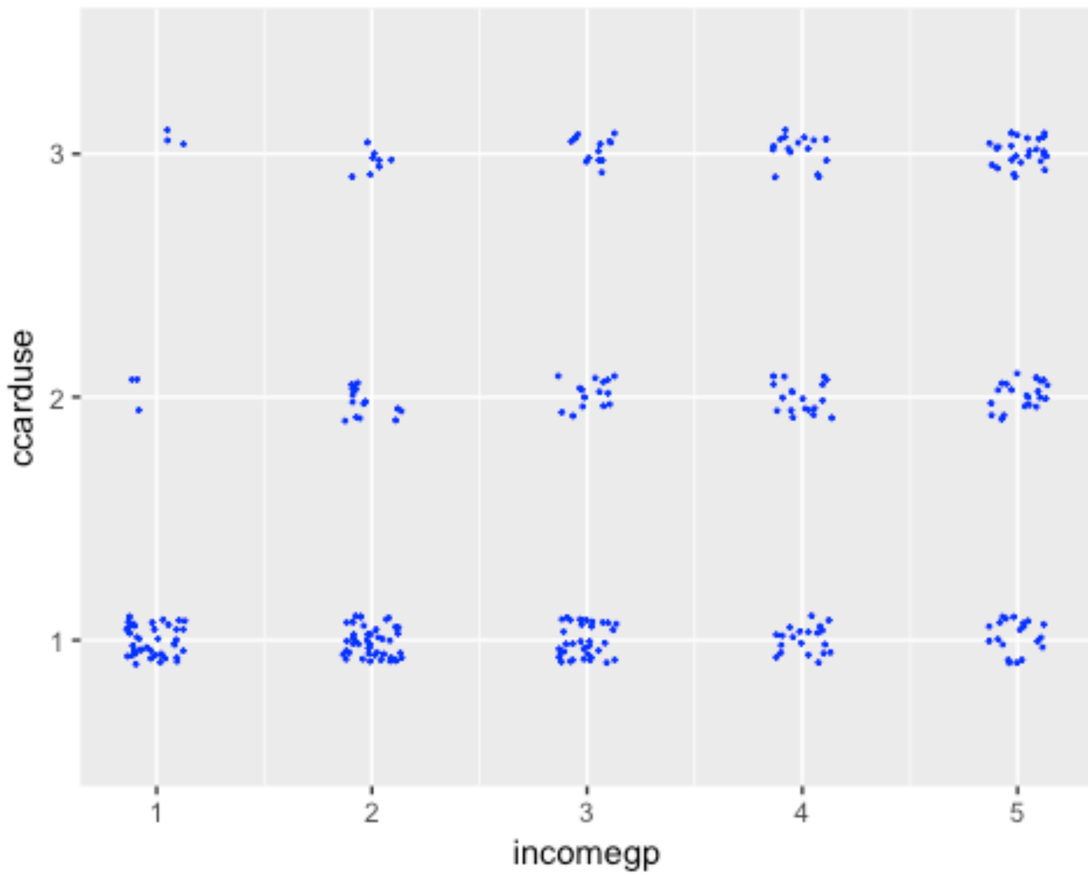
```
library(faraway)
data("debt")

# Check for NA's in the data and omit them.
debt <- na.omit(debt)
summary(debt)
```

```
##     incomegp         house          children         singpar
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :3.000   Median :2.000   Median :1.0000   Median :0.00000
##  Mean   :3.105   Mean   :2.043   Mean   :0.9605   Mean   :0.05592
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:2.0000   3rd Qu.:0.00000
##  Max.   :5.000   Max.   :3.000   Max.   :4.0000   Max.   :1.00000
```

```
##      agegp          bankacc          bsocacc          manage     
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :1.000  
##  1st Qu.:2.000   1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:4.000  
##  Median :2.000   Median :1.0000   Median :1.000   Median :4.000  
##  Mean   :2.461   Mean   :0.8421   Mean   :0.625   Mean   :4.207  
##  3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:5.000  
##  Max.   :4.000   Max.   :1.0000   Max.   :1.000   Max.   :5.000  
##     ccarduse         cigbuy          xmasbuy          locintrn    
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :1.500  
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:3.830  
##  Median :1.000   Median :0.0000   Median :1.000   Median :4.415  
##  Mean   :1.701   Mean   :0.3191   Mean   :0.875   Mean   :4.413  
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:5.000  
##  Max.   :3.000   Max.   :1.0000   Max.   :1.000   Max.   :7.000  
##     prodebt     
##  Min.   :1.350  
##  1st Qu.:2.710  
##  Median :3.180  
##  Mean   :3.199  
##  3rd Qu.:3.650  
##  Max.   :5.470  

# 5.a) Declare the response as an ordered factor and make a plot showing the
relationship to prodebt.

# Declare the response variable ccarduse as ordered.
debt$ccarduse <- factor(debt$ccarduse, ordered = TRUE)

# Verify that the response is indeed ordered.
is.ordered(debt$ccarduse)

## [1] TRUE

# Create the plot
library(ggplot2)
ggplot(debt, aes(x = prodebt, y = ccarduse)) +
  geom_jitter(width = 0, height = 0.1, size = 0.15, color = "blue")
```




```r
# In this plot we can see a more pronounced relationship between the lowest frequency of credit card use, i.e., never, and debt attitude, particularly with lower quantities of the debt attitude score, i.e., less favorable to debt,  However, for higher levels of frequency of card use, 2 and 3, the data points seem  to be similarly scattered, slightly denser in the mid-range of the debt attitude scores.

# Then response against the income group.
ggplot(debt, aes(x = incomegp, y = ccarduse)) +
  geom_jitter(width = 0.14, height = 0.1, size = 0.24, color = "blue")
```




```
# In this plot as well, a more pronounced relationship can be seen between
the lowest level  of the card use frequency, and income group, particularly
the lower levels of income group.  Another slightly pronounced relationship
between the third level of card use frequency and  the fifth level of the
income group is visible.


# 5.b) Fit a Proportional Odds model for credit card use to all other
variables.

library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

# Fit the Proportional Odds model
pomod <- polr(ccarduse ~ ., debt )
# Look at the coefficients for largest t-values.
summary(pomod)$coef
```

```
##
## Re-fitting to get Hessian

##                Value Std. Error    t value
## incomegp  0.47131302  0.1060967  4.4422968
## house     0.11600148  0.2323630  0.4992251
## children -0.07872411  0.1250325 -0.6296291
## singpar   0.88171828  0.5971140  1.4766330
## agegp     0.20568368  0.1576103  1.3050145
## bankacc   2.10269577  0.5933918  3.5435203
## bsocacc   0.47321630  0.2671328  1.7714643
## manage    0.18179169  0.1652902  1.0998331
## cigbuy   -0.73545858  0.2980681 -2.4674178
## xmasbuy   0.47014289  0.4129631  1.1384622
## locintrn  0.11881236  0.1423979  0.8343685
## prodebt   0.61046374  0.1822466  3.3496579
## 1|2       7.96937466  1.4751711  5.4023391
## 2|3       9.39436162  1.5051079  6.2416532

# The two most significant predictors (having largest t-values) are the
"incomegp" and "bankacc", the later having a t-value that is close to that of
the "prodebt" variable.

# bankacc is obviously significantly influential on ccarduse since it is very
unlikely to  have a credit card without a bank account. The coefficient for
bankacc is about 2.1 which  means that the the odds of moving from ccarduse
level of 1 to 2/3 or from 1/2 to 3 increase  by a factor of exp(2.1) = 8.16
when the bankacc is equal to 1. The coefficient for incomegp is about 0.47.
That is, the odds of moving from ccarduse level of 1 to 2/3 or from 1/2 to 3
increase by a factor of exp(0.47) = 1.60 when income goes to next level.

# The least significant predictor is security of housing tenure with a t-
value of 0.499.

# 5.c) Fit a Proportional Odds model using the least significant predictor.

pomodhouse <- polr(ccarduse ~ house, debt)
summary(pomodhouse)$coef

##
## Re-fitting to get Hessian

##           Value Std. Error  t value
## house 0.5626874  0.1770438 3.178238
## 1|2   1.2821108  0.3870200 3.312776
## 2|3   2.3980930  0.4032797 5.946476

# We see that the t-value for the variable "house" in this model is
significantly larger  than in the full model with a lower standard error. The
t-value is close to that of  "bankacc" and "prodebt" in the full model, so it
is much more significant than the full model suggests.
```

```r
# 5.d) Use stepwise AIC to reduce the full model.

reducedpomod <- step(pomod, trace = 0)

summary(reducedpomod)

##
## Re-fitting to get Hessian

## Call:
## polr(formula = ccarduse ~ incomegp + agegp + bankacc + bsocacc +
##     cigbuy + prodebt, data = debt)
##
## Coefficients:
##            Value Std. Error t value
## incomegp  0.4589     0.1007   4.555
## agegp     0.2696     0.1352   1.993
## bankacc   2.0816     0.5753   3.618
## bsocacc   0.5048     0.2591   1.949
## cigbuy   -0.7677     0.2922  -2.627
## prodebt   0.5635     0.1755   3.211
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  5.9944  0.9961     6.0178
## 2|3  7.3948  1.0276     7.1961
##
## Residual Deviance: 517.5895
## AIC: 533.5895

# The qualitative effect of the predictor, as discussed above, is that the
odds of moving to the next level of ccarduse increases by a factor of
exp(coefficient) when the corresponding predictor is increased by one unit or
moves to the nect level. We see that the predictor "house" is dropped from
the model, but we saw above that this predictor is significant when we use
only that in the one-predictor model.

# 5.e) Compute the median value of the predictors in the reduced model.

median.inc <- median(debt$incomegp)
median.age <- median(debt$agegp)
median.bank <- median(debt$bankacc)
median.bsoc <- median(debt$bsocacc)
median.cig <- median(debt$cigbuy)
median.pdebt <- median(debt$prodebt)

# Compute the predicted probabilities at the median values for smokers
predict(reducedpomod, data.frame(incomegp = median.inc, agegp = median.age,
```

```r
bankacc = median.bank, bsocacc = median.bsoc,cigbuy = debt$cigbuy[debt$cig[]
== 1], prodebt = median.pdebt), type = "probs")[1,]

##         1         2         3
## 0.6149076 0.2513666 0.1337258

# Compute the predicted probabilities at the median values for non-smokers
predict(reducedpomod, data.frame(incomegp = median.inc, agegp = median.age,
bankacc = median.bank, bsocacc = median.bsoc, cigbuy = debt$cigbuy[debt$cig[]
== 0], prodebt = median.pdebt), type = "probs")[1,]

##         1         2         3
## 0.4256250 0.3247658 0.2496092

# The highest probability in both cases is for the first level of ccarduse.So
both groups have a higher probaility of never using their ccards and a lower
probability of using their cards regularly, while non-smokers have a higher
probability of regularly using their cards than smokers do.

# 5.f) Fit a Proportional Hazards model

phmod <- polr(ccarduse ~ incomegp + agegp + bankacc + bsocacc +
                cigbuy + prodebt, data = debt, method = "cloglog")
summary(phmod)

##
## Re-fitting to get Hessian

## Call:
## polr(formula = ccarduse ~ incomegp + agegp + bankacc + bsocacc +
##     cigbuy + prodebt, data = debt, method = "cloglog")
##
## Coefficients:
##            Value Std. Error t value
## incomegp  0.2454    0.05950   4.125
## agegp     0.1936    0.08224   2.354
## bankacc   0.9984    0.23658   4.220
## bsocacc   0.3087    0.15704   1.966
## cigbuy   -0.3120    0.15789  -1.976
## prodebt   0.3418    0.10872   3.143
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  3.0002  0.5307     5.6536
## 2|3  3.8261  0.5424     7.0541
##
## Residual Deviance: 527.372
## AIC: 543.372

# Recompute the two sets of probabilities from the previous part.
```

```r
predict(phmod, data.frame(incomegp = median.inc, agegp = median.age,
                                 bankacc = median.bank, bsocacc =
median.bsoc,
                                 cigbuy = debt$cigbuy[debt$cig[] == 1],
                                 prodebt = median.pdebt), type = "probs")[1,]

##         1         2         3
## 0.5571469 0.2872181 0.1556350

predict(phmod, data.frame(incomegp = median.inc, agegp = median.age,
                                 bankacc = median.bank, bsocacc =
median.bsoc,
                                 cigbuy = debt$cigbuy[debt$cig[] == 0],
                                 prodebt = median.pdebt), type = "probs")[1,]

##         1         2         3
## 0.4491074 0.2946605 0.2562321

# Using the Proportional Hazards model seems to make the two sets of
probailities more similar to eachother, making it almost equally likely for
both smokers and nonsmokers to never, often, or regularly use their
creditcards. The non-smokeres,however, are still more likely to use their
credit cars regularly than are smokers.
```