NCSU ST 503 Discussion 13

Probem 7.1 Faraway, Julian J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models CRC Press.

Bruce Campbell

7.1

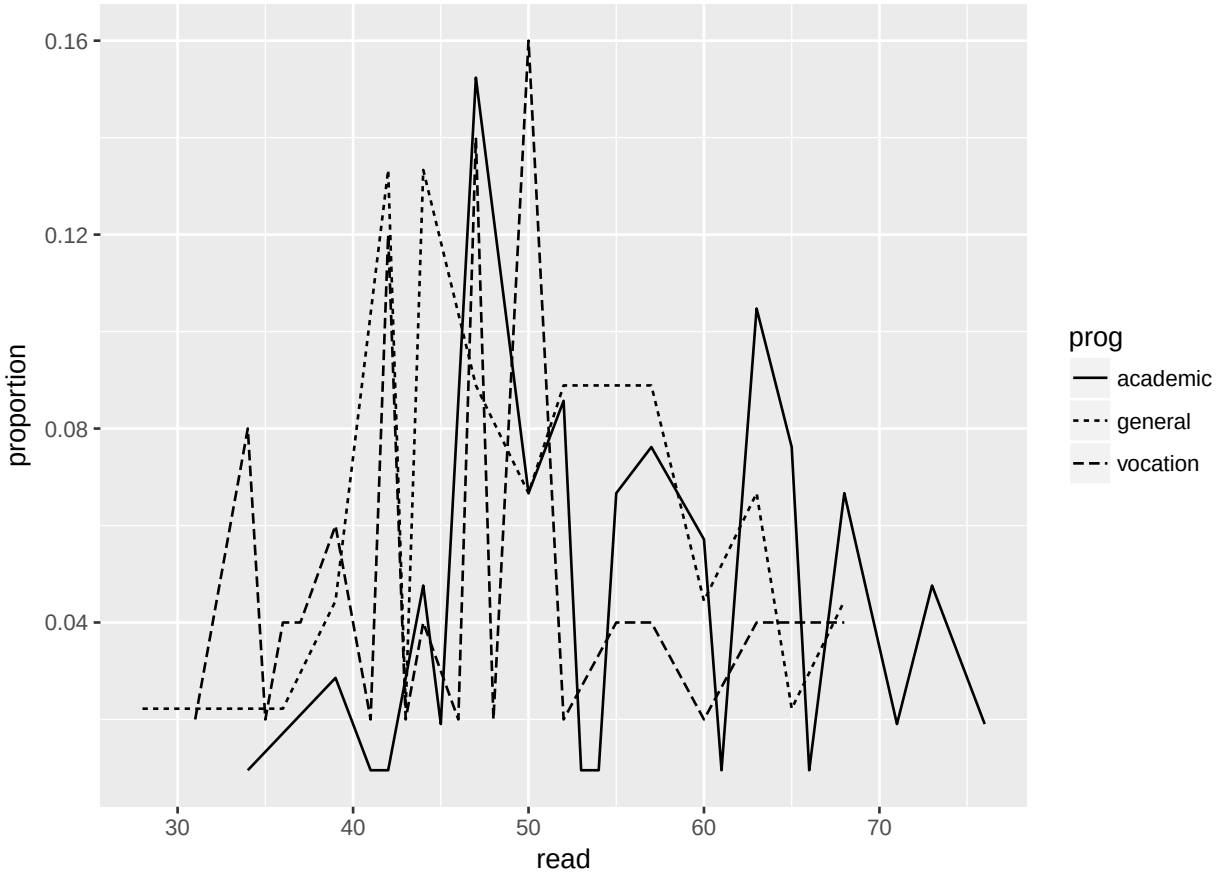
The hsb data was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The variables are gender; race; socioeconomic status (SES); school type; chosen high school program type; scores on reading, writing, math, science, and social studies. We want to determine which factors are related to the choice of the type of program - academic, vocational or general - that the students pursue in high school. The response is multinomial with three levels.

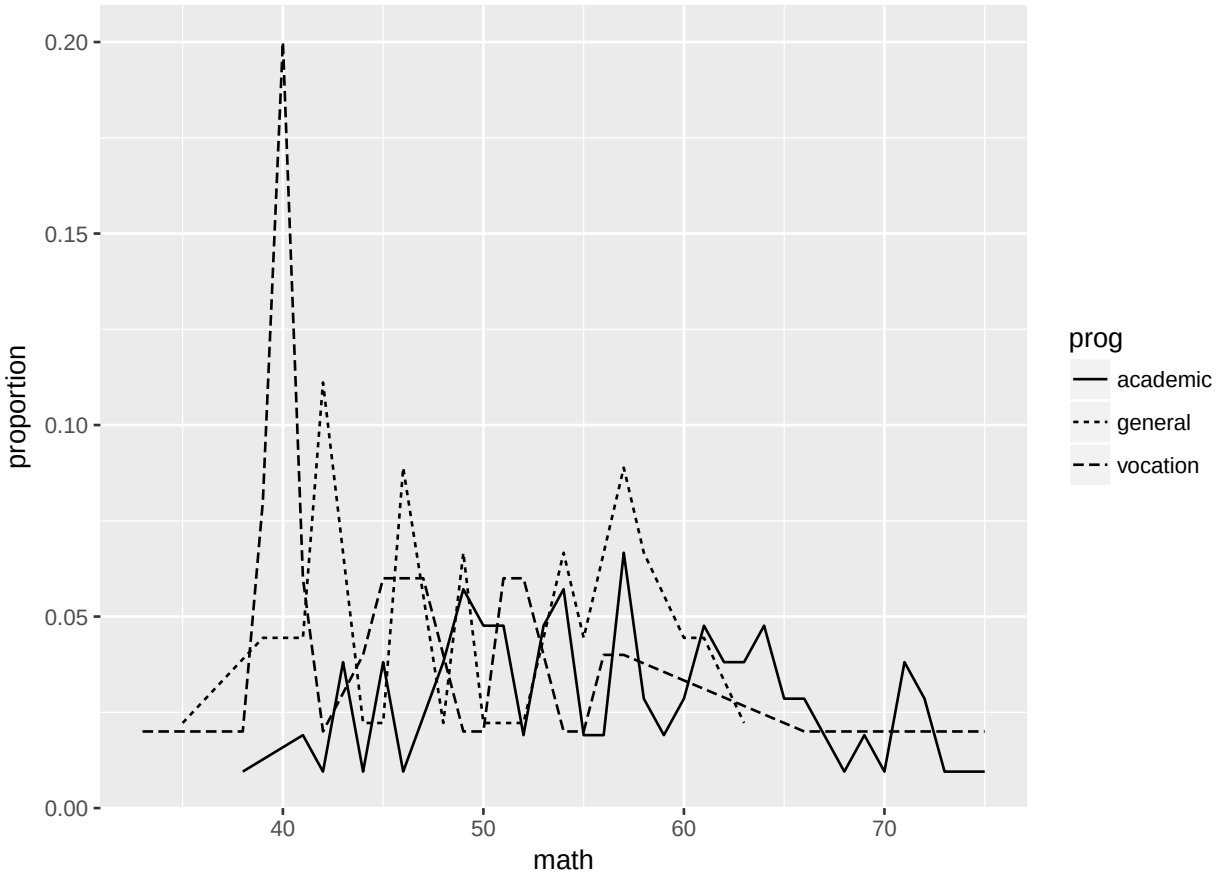
(a) Make a table showing the proportion of males and females choosing the three different programs. Comment on the difference. Repeat this comparison but for SES rather than gender.

```
##           gender
## prog       female male
##   academic     58   47
##   general      24   21
##   vocation     27   23
```

```
##           ses
## prog       high low middle
##   academic   42  19     44
##   general     9  16     20
##   vocation    7  12     31
```

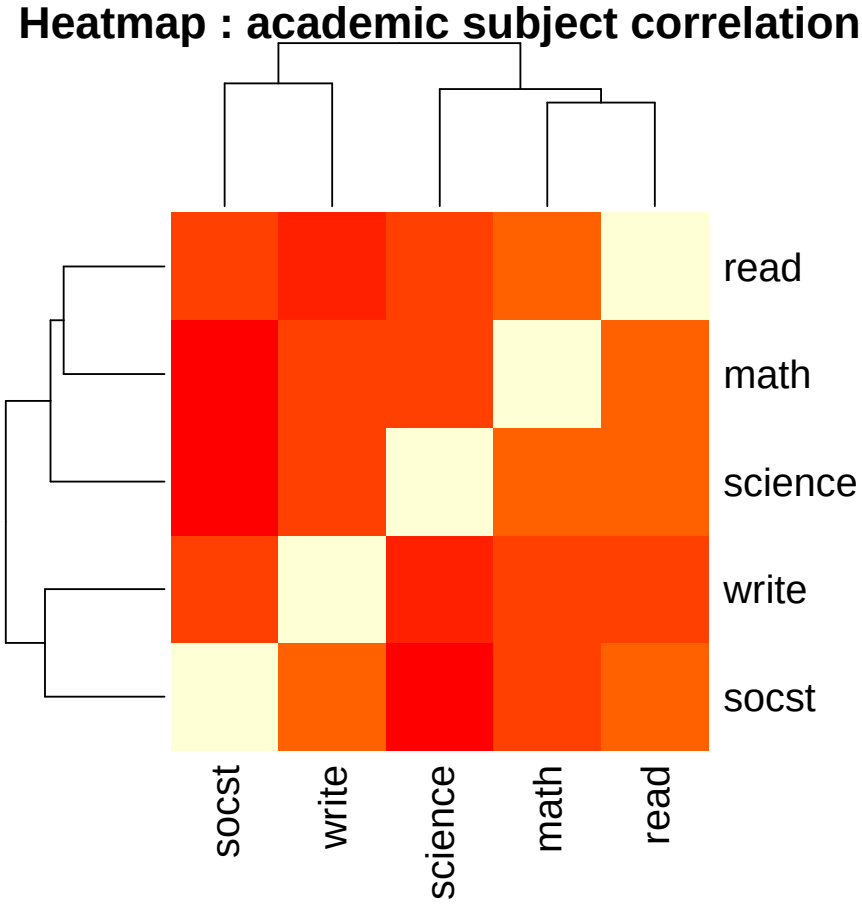
(b) Construct a plot like the right panel of Figure 7.1 that shows the relationship between program choice and reading score. Comment on the plot. Repeat for math in place of reading.





(c) Compute the correlation matrix for the five subject scores.

	read	write	math	science	socst
read	1	0.5968	0.6623	0.6302	0.6215
write	0.5968	1	0.6174	0.5704	0.6048
math	0.6623	0.6174	1	0.6307	0.5445
science	0.6302	0.5704	0.6307	1	0.4651
socst	0.6215	0.6048	0.5445	0.4651	1



(d) Fit a multinomial response model for the program choice and examine the fitted coefficients. Of the five subjects, one gives unexpected coefficients. Identify this subject and suggest an explanation for this behavior.

```
## # 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
```

```
## Call:
## multinom(formula = prog ~ ., data = df)
##
## 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
```

(e) Construct a derived variable that is the sum of the five subject scores. Fit a multinomial model as before except with this one sum variable in place of the five subjects separately. Compare the two models to decide which should be preferred.

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

```
## Call:
## multinom(formula = prog ~ id + gender + race + ses + schtyp +
##     sum.subject, data = df.reduced)
##
## 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 sum.subject
## 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 sum.subject
## 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
```

The s.e. for the combined subject variable is much lower than the single subjet variables. We suspect colliniearity may be the cause.

(f) Use a stepwise method to reduce the model. Which variables are in your selected model?

```
## Call:
## multinom(formula = prog ~ ses + schtyp + sum.subject, data = df.reduced)
##
## Coefficients:
##          (Intercept)    seslow sesmiddle schtyppublic sum.subject
## general     2.593944 0.8078324 0.5808536    0.5594952 -0.01635887
## vocation    6.372051 0.1330839 1.1517240    1.8490860 -0.03681150
##
## Std. Errors:
##          (Intercept)    seslow sesmiddle schtyppublic sum.subject
## general     1.587502 0.5386033 0.4720925    0.5219044 0.005422494
## vocation    1.877764 0.6468558 0.5465572    0.7974692 0.006553295
##
## Residual Deviance: 336.0554
## AIC: 356.0554
```

We see that there are 3 variables in the best model : $ses + schtyp + sum.subjec$

(g) Construct a plot of predicted probabilities from your selected model where the math score varies over the observed range. Other predictors should be set at the most common level or mean value as appropriate. Your plot should be similar to Figure 7.2. Comment on the relationship.