

STAT-S632

Assignment 5

John Koo

```
import::from(magrittr, `>`, `<>`)
library(ggplot2)
import::from(nnet, multinom)

theme_set(theme_bw())
```

Problem 1

Part a

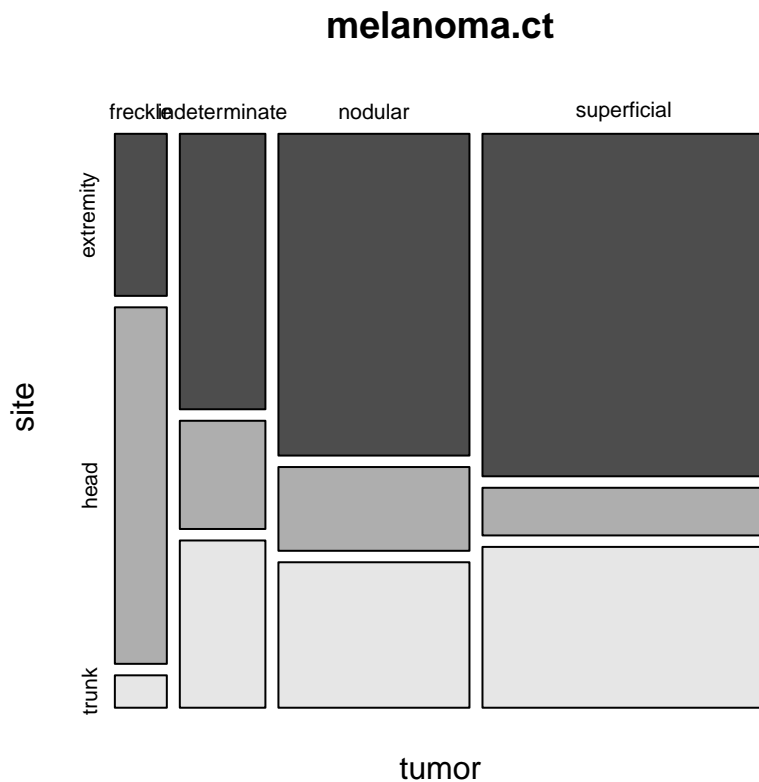
```
# get the data
melanoma.df <- faraway::melanoma
summary(melanoma.df)
```

	count	tumor	site
Min. :	2.00	freckle	:3 extremity:4
1st Qu.:	14.75	indeterminate:	3 head :4
Median :	20.50	nodular	:3 trunk :4
Mean :	33.33	superficial	:3
3rd Qu.:	38.25		
Max. :	115.00		

```
# contingency table stuff
melanoma.ct <- xtabs(count ~ tumor + site, data = melanoma.df)
melanoma.ct
```

		site		
tumor		extremity	head	trunk
freckle		10	22	2
indeterminate		28	11	17
nodular		73	19	33
superficial		115	16	54

```
mosaicplot(melanoma.ct, color = TRUE)
```



```
# poisson model
pois.mod <- glm(count ~ tumor + site, data = melanoma.df, family = poisson)
summary(pois.mod)
```

Call:

```
glm(formula = count ~ tumor + site, family = poisson, data = melanoma.df)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.0453	-1.0741	0.1297	0.5857	5.1354

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.9554	0.1770	16.696	< 2e-16 ***
tumorindeterminate	0.4990	0.2174	2.295	0.0217 *
tumornodular	1.3020	0.1934	6.731	1.68e-11 ***
tumorsuperficial	1.6940	0.1866	9.079	< 2e-16 ***
sitehead	-1.2010	0.1383	-8.683	< 2e-16 ***
sitetrunk	-0.7571	0.1177	-6.431	1.27e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 295.203 on 11 degrees of freedom
Residual deviance: 51.795 on 6 degrees of freedom
AIC: 122.91
```

Number of Fisher Scoring iterations: 5

```
pchisq(pois.mod$deviance, pois.mod$df.residual, lower.tail = FALSE)
```

```
[1] 2.050453e-09
```

The Poisson model is not a good fit according to a χ^2 test on the deviance of the model, and this could be due to the fact that the regressors `tumor` and `site` are not independent. We can see some of this in the mosaic plot—freckles are much more common on heads compared to other types of tumors.

Part b

```
melanoma.df %<>% dplyr::mutate(pois.resid = residuals(pois.mod))
resid.ct <- xtabs(pois.resid ~ tumor + site, data = melanoma.df)
resid.ct
```

	site		
tumor	extremity	head	trunk
freckle	-2.31583297	5.13537787	-2.82829426
indeterminate	-0.66016102	0.46798432	0.54787007
nodular	0.28104581	-0.49711084	-0.02173229
superficial	1.00813975	-3.04533605	0.69899703

We see large residuals for the tumor type “freckle”.

Problem 2

Part a

```
uncviet.df <- faraway::uncviet
summary(uncviet.df)
```

	y	policy	sex	year
Min.	: 3.00	A:10	Female:20	Fresh :8
1st Qu.:	18.50	B:10	Male :20	Grad :8
Median :	42.00	C:10		Junior:8
Mean :	78.67	D:10		Senior:8
3rd Qu.:	131.25			Soph :8
Max.	:345.00			

```
uncviet.ct <- xtabs(y ~ policy + sex + year, data = uncviet.df)
uncviet.ct
```

```
, , year = Fresh
```

```
      sex
policy Female Male
A         13  175
```

B	19	116
C	40	131
D	5	17

, , year = Grad

	sex	
policy	Female	Male
A	19	118
B	27	176
C	128	345
D	13	141

, , year = Junior

	sex	
policy	Female	Male
A	22	132
B	29	120
C	110	154
D	6	29

, , year = Senior

	sex	
policy	Female	Male
A	12	145
B	21	95
C	58	185
D	10	44

, , year = Soph

	sex	
policy	Female	Male
A	5	160
B	9	126
C	33	135
D	3	21

```
summary(uncviet.ct)
```

Call: xtabs(formula = y ~ policy + sex + year, data = uncviet.df)

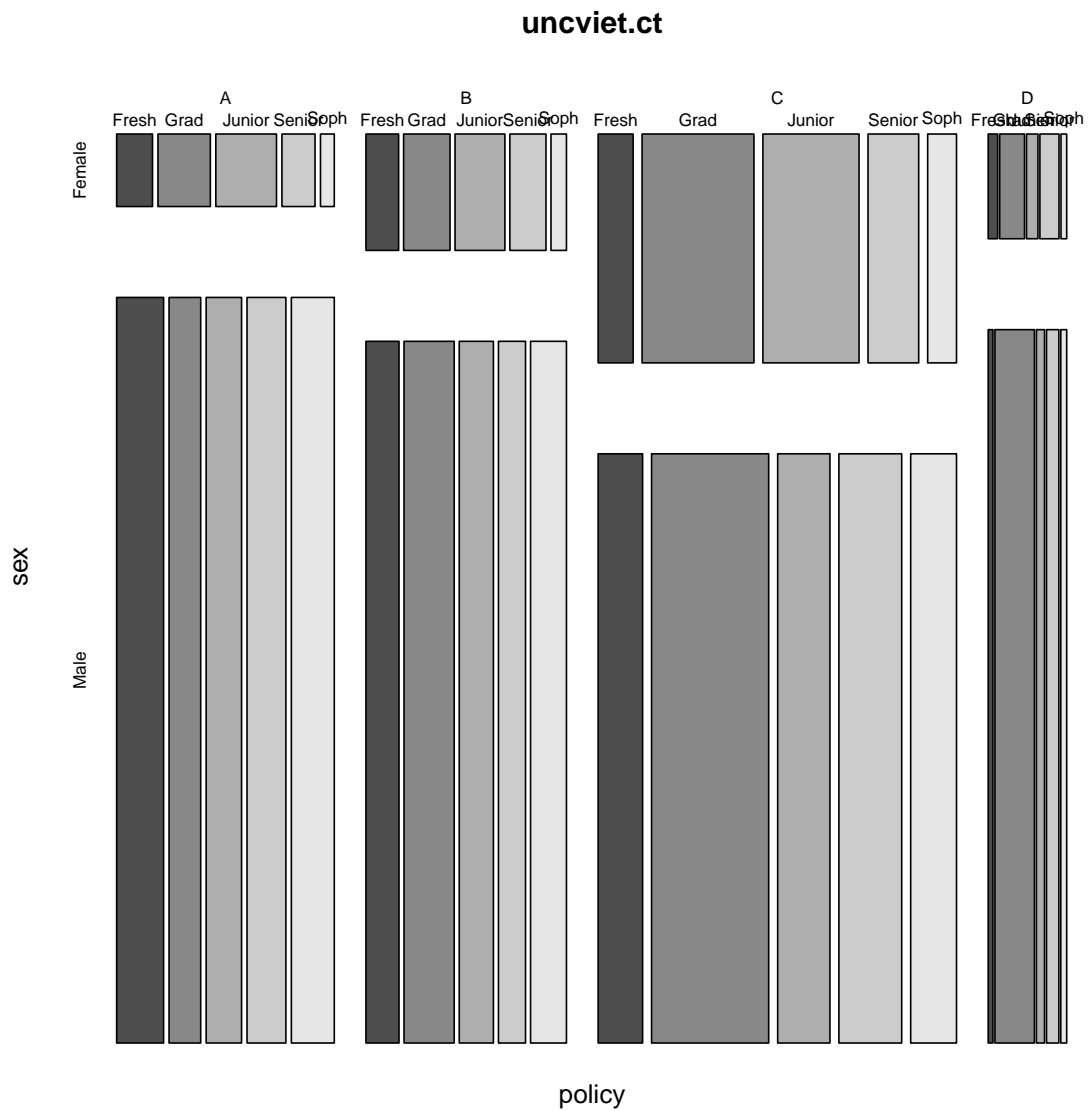
Number of cases in table: 3147

Number of factors: 3

Test for independence of all factors:

Chisq = 449.2, df = 31, p-value = 1.118e-75

```
mosaicplot(uncviet.ct, color = TRUE)
```



```
mantelhaen.test(uncviet.ct)
```

Cochran-Mantel-Haenszel test

```
data: uncviet.ct
```

```
Cochran-Mantel-Haenszel M^2 = 132.55, df = 3, p-value < 2.2e-16
```

```
uncviet.pois.mod <- glm(y ~ policy + sex + year, data = uncviet.df,
                        family = poisson)
```

```
summary(uncviet.pois.mod)
```

Call:

```
glm(formula = y ~ policy + sex + year, family = poisson, data = uncviet.df)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.323	-2.582	-0.810	0.673	7.873

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.19003	0.06534	48.823	<2e-16 ***
policyB	-0.08192	0.05102	-1.605	0.1084
policyC	0.49877	0.04479	11.134	<2e-16 ***
policyD	-1.01943	0.06862	-14.856	<2e-16 ***
sexMale	1.48324	0.04591	32.305	<2e-16 ***
yearGrad	0.62809	0.05452	11.521	<2e-16 ***
yearJunior	0.15415	0.05999	2.569	0.0102 *
yearSenior	0.09953	0.06076	1.638	0.1014
yearSoph	-0.04763	0.06301	-0.756	0.4497

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2708.08 on 39 degrees of freedom
Residual deviance: 423.83 on 31 degrees of freedom
AIC: 666.22

Number of Fisher Scoring iterations: 5

```
pchisq(uncviet.pois.mod$deviance, uncviert.pois.mod$df.residual,  
       lower.tail = FALSE)
```

```
[1] 1.591439e-70
```

We have evidence from the mosaic plot, contingency tables, Mantel-Haenszel test, and Poisson model that the three regressors are not independent. So we might be interested in seeing how they are dependent. Visually, we can see some relationship between sex and policy. We can also see some relationship between year and policy as well. Instead of considering the different possibilities, we can just use AIC:

```
uncviet.nominal.mod <- glm(y ~ (policy + sex + year) ** 2, data = uncviert.df,  
                           family = poisson) %>%  
  step(direction = 'both')
```

Start: AIC=299.58

y ~ (policy + sex + year)^2

	Df	Deviance	AIC
<none>		19.194	299.58
- sex:year	4	70.643	343.03
- policy:sex	3	153.935	428.32
- policy:year	12	216.312	472.70

```
summary(uncviet.nominal.mod)
```

Call:

```
glm(formula = y ~ (policy + sex + year)^2, family = poisson,  
     data = uncviert.df)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4849	-0.4420	0.0023	0.3962	1.8756

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.69824	0.16895	15.970	< 2e-16 ***
policyB	0.13458	0.18626	0.723	0.469945
policyC	1.05527	0.16400	6.435	1.24e-10 ***
policyD	-1.78914	0.29955	-5.973	2.33e-09 ***
sexMale	2.45589	0.16548	14.841	< 2e-16 ***
yearGrad	-0.21730	0.18001	-1.207	0.227367
yearJunior	0.40703	0.17827	2.283	0.022415 *
yearSenior	-0.09396	0.18938	-0.496	0.619800
yearSoph	-0.57514	0.21360	-2.693	0.007090 **
policyB:sexMale	-0.51798	0.16448	-3.149	0.001637 **
policyC:sexMale	-1.35481	0.14116	-9.598	< 2e-16 ***
policyD:sexMale	-0.39366	0.21999	-1.789	0.073545 .
policyB:yearGrad	0.71910	0.15814	4.547	5.44e-06 ***
policyC:yearGrad	1.31478	0.14590	9.011	< 2e-16 ***
policyD:yearGrad	2.25855	0.25416	8.886	< 2e-16 ***
policyB:yearJunior	0.25659	0.16181	1.586	0.112799
policyC:yearJunior	0.49017	0.15030	3.261	0.001109 **
policyD:yearJunior	0.63363	0.29360	2.158	0.030915 *
policyB:yearSenior	0.02394	0.16669	0.144	0.885809
policyC:yearSenior	0.51505	0.15011	3.431	0.000601 ***
policyD:yearSenior	1.07484	0.27513	3.907	9.36e-05 ***
policyB:yearSoph	0.14893	0.16213	0.919	0.358320
policyC:yearSoph	0.18157	0.15499	1.171	0.241406
policyD:yearSoph	0.23077	0.31400	0.735	0.462372
sexMale:yearGrad	-0.10814	0.15328	-0.706	0.480479
sexMale:yearJunior	-0.68075	0.15747	-4.323	1.54e-05 ***
sexMale:yearSenior	-0.09400	0.16944	-0.555	0.579056
sexMale:yearSoph	0.47497	0.19722	2.408	0.016024 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2708.080 on 39 degrees of freedom
Residual deviance: 19.194 on 12 degrees of freedom
AIC: 299.58

Number of Fisher Scoring iterations: 4

```
pchisq(uncviet.nominal.mod$deviance,  
       uncviet.nominal.mod$df.residual,  
       lower.tail = FALSE)
```

[1] 0.08394884

It appears that the full model (with two-way interactions) provides the lowest AIC, and a χ^2 test suggests that this may be a decent fit. This implies that there is full dependence (including pairwise dependence between all three pairs of regressors).

Part b

```
uncviet.df %<>%
  dplyr::mutate(policy.ord = as.numeric(factor(policy, levels = LETTERS[1:4])),
               year.ord = as.numeric(factor(year, levels = c('Fresh',
                                                           'Soph',
                                                           'Junior',
                                                           'Senior',
                                                           'Grad'))))
uncviet.ord.mod <- glm(y ~ policy + sex + year + I(policy.ord * year.ord),
                     data = uncviet.df, family = poisson)
anova(uncviet.pois.mod, uncviet.ord.mod, test = 'Chi')
```

Analysis of Deviance Table

Model 1: y ~ policy + sex + year

Model 2: y ~ policy + sex + year + I(policy.ord * year.ord)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	31	423.83			
2	30	246.13	1	177.69	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
summary(uncviet.ord.mod)$coef['I(policy.ord * year.ord)', ]
```

	Estimate	Std. Error	z value	Pr(> z)
1	1.751092e-01	1.354426e-02	1.292867e+01	3.101548e-38

We have evidence of association from both the χ^2 test and the Wald test. The estimate $\hat{\gamma} > 0$, suggesting positive association. This suggests that as **year** increases, there is a higher probability that the responder favors less involvement in the war.

Problem 3

We are given:

$$\eta_{ij} = \log \frac{p_{ij}}{p_{i1}}$$

Then we get:

$$\begin{aligned} e^{\eta_{ij}} &= \frac{p_{ij}}{p_{i1}} \\ p_{i1} e^{\eta_{ij}} &= p_{ij} \\ p_{i1} \sum_j e^{\eta_{ij}} &= \sum_j p_{ij} = 1 \end{aligned}$$

We can also say that $\sum_j e^{\eta_{ij}} = \sum_{j=2} e^{\eta_{ij}} + e^{\eta_{i1}}$

We can also see that $\eta_{i1} = \log \frac{p_{i1}}{p_{i1}} = \log 1 = 0$, so $e^{\eta_{i1}} = 1$.

And finally, we can see that since $\eta_{ij} = \log \frac{p_{ij}}{p_{i1}}$, $p_{i1} = \frac{p_{ij}}{e^{\eta_{ij}}}$. Putting this all together:

$$p_{i1} \sum_j e^{\eta_{ij}} = 1$$

$$\frac{p_{ij}}{e^{\eta_{ij}}} (1 + \sum_{j=2}^J e^{\eta_{ij}}) = 1$$

$$p_{ij} = \frac{e^{\eta_{ij}}}{1 + \sum_{j=2}^J e^{\eta_{ij}}}$$

Problem 4

Part a

```
hsb.df <- faraway::hsb %>%
  dplyr::mutate(ses = factor(ses, levels = c('low', 'middle', 'high')))
summary(hsb.df)
```

id	gender	race	ses
Min. : 1.00	female:109	african-amer: 20	low :47
1st Qu.: 50.75	male : 91	asian : 11	middle:95
Median :100.50		hispanic : 24	high :58
Mean :100.50		white :145	
3rd Qu.:150.25			
Max. :200.00			

schtyp	prog	read	write
private: 32	academic:105	Min. :28.00	Min. :31.00
public :168	general : 45	1st Qu.:44.00	1st Qu.:45.75
	vocation: 50	Median :50.00	Median :54.00
		Mean :52.23	Mean :52.77
		3rd Qu.:60.00	3rd Qu.:60.00
		Max. :76.00	Max. :67.00

math	science	socst
Min. :33.00	Min. :26.00	Min. :26.00
1st Qu.:45.00	1st Qu.:44.00	1st Qu.:46.00
Median :52.00	Median :53.00	Median :52.00
Mean :52.65	Mean :51.85	Mean :52.41
3rd Qu.:59.00	3rd Qu.:58.00	3rd Qu.:61.00
Max. :75.00	Max. :74.00	Max. :71.00

```
hsb.df %>%
  dplyr::group_by(prog, gender) %>%
  dplyr::summarise(y = n()) %>%
  dplyr::ungroup() %>%
  xtabs(y ~ gender + prog, data = .) %>%
  prop.table(1)
```

	academic	general	vocation
female	0.5321101	0.2201835	0.2477064
male	0.5164835	0.2307692	0.2527473

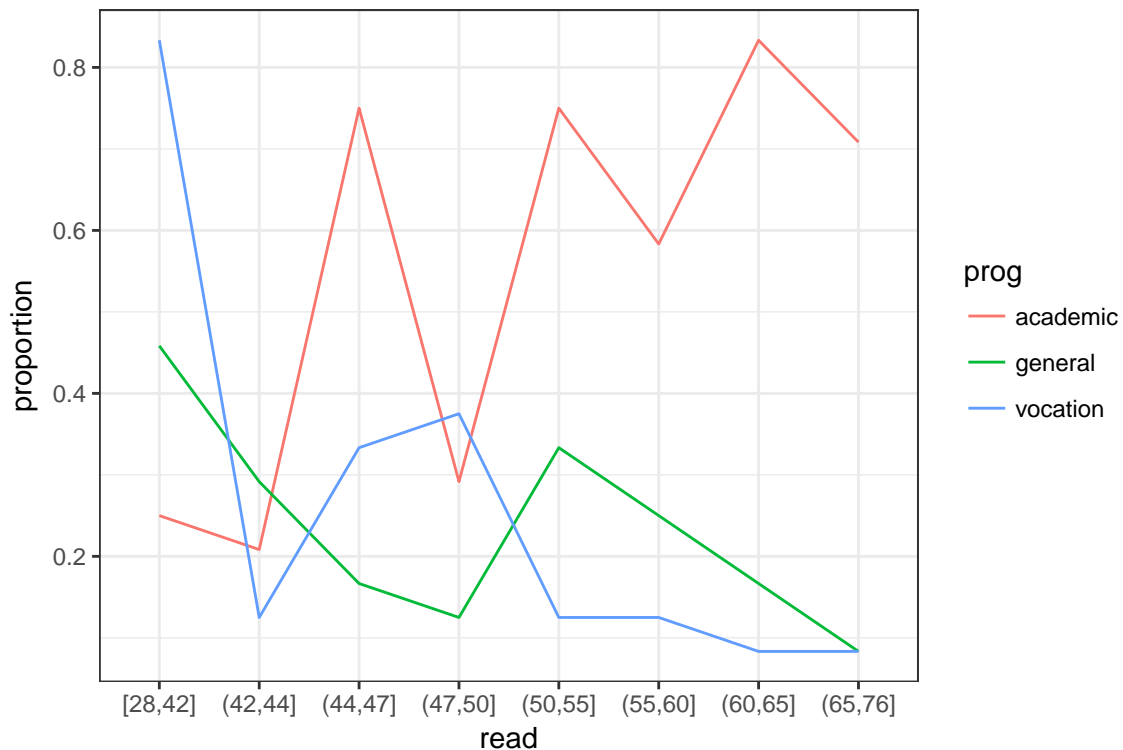
```
hsb.df %>%
  dplyr::group_by(prog, ses) %>%
  dplyr::summarise(y = n()) %>%
  dplyr::ungroup() %>%
  xtabs(y ~ ses + prog, data = .) %>%
  prop.table(1)
```

	prog		
ses	academic	general	vocation
low	0.4042553	0.3404255	0.2553191
middle	0.4631579	0.2105263	0.3263158
high	0.7241379	0.1551724	0.1206897

It appears that there is no strong association between **gender** and **prog**, but there is one between **ses** and **prog**.

Part b

```
hsb.df %>%
  dplyr::mutate(read = cut_number(read, 8)) %>%
  dplyr::group_by(prog, read) %>%
  dplyr::summarise(y = n()) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(proportion = y / n()) %>%
  ggplot() +
  geom_line(aes(x = read, y = proportion, group = prog, colour = prog))
```



It appears that as reading scores increase, the probability of being in an academic program increases, and as

reading scores decrease, the probability of being in a vocational program increases.

Part c

```
hsb.df %<>%  
  dplyr::mutate(ses = as.numeric(ses)) # to ordinal  
prog.mod <- multinom(prog ~ ., data = hsb.df) %>%  
  step(direction = 'both', trace = FALSE)
```

```
# weights: 42 (26 variable)  
initial value 219.722458  
iter 10 value 186.536640  
iter 20 value 158.410762  
iter 30 value 156.034011  
final value 156.033991  
converged  
trying - id  
trying - gender  
trying - race  
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 186.710606  
iter 20 value 161.060816  
final value 159.177698  
converged  
trying - id  
trying - gender  
trying - ses  
trying - schtyp  
trying - read  
trying - write  
trying - math  
trying - science  
trying - socst  
trying + race  
# weights: 30 (18 variable)  
initial value 219.722458  
iter 10 value 186.000556  
iter 20 value 159.934297  
final value 159.409406  
converged  
trying - id  
trying - ses  
trying - schtyp  
trying - read  
trying - write
```

```

trying - math
trying - science
trying - socst
trying + gender
trying + race
# weights: 27 (16 variable)
initial value 219.722458
iter 10 value 179.532582
iter 20 value 159.734855
final value 159.730690
converged
trying - ses
trying - schtyp
trying - read
trying - write
trying - math
trying - science
trying - socst
trying + id
trying + gender
trying + race
# weights: 24 (14 variable)
initial value 219.722458
iter 10 value 181.921322
iter 20 value 160.333721
final value 160.333655
converged
trying - ses
trying - schtyp
trying - read
trying - math
trying - science
trying - socst
trying + id
trying + gender
trying + race
trying + write
# weights: 21 (12 variable)
initial value 219.722458
iter 10 value 174.300997
iter 20 value 161.857206
iter 20 value 161.857205
iter 20 value 161.857205
final value 161.857205
converged
trying - ses
trying - schtyp
trying - math
trying - science
trying - socst
trying + id
trying + gender
trying + race
trying + read

```

```

trying + write
# weights:  18 (10 variable)
initial value 219.722458
iter  10 value 166.035430
final value 163.818866
converged
trying - schtyp
trying - math
trying - science
trying - socst
trying + id
trying + gender
trying + race
trying + ses
trying + read
trying + write

```

```
summary(prog.mod)
```

Call:

```
multinom(formula = prog ~ schtyp + math + science + socst, data = hsb.df)
```

Coefficients:

	(Intercept)	schtyppublic	math	science	socst
general	3.854099	0.6735847	-0.1205511	0.07441108	-0.05144098
vocation	7.022897	1.7880022	-0.1370433	0.04214340	-0.08672034

Std. Errors:

	(Intercept)	schtyppublic	math	science	socst
general	1.512678	0.5336615	0.03182842	0.02727090	0.02314651
vocation	1.739585	0.8041796	0.03545611	0.02773962	0.02471897

Residual Deviance: 327.6377

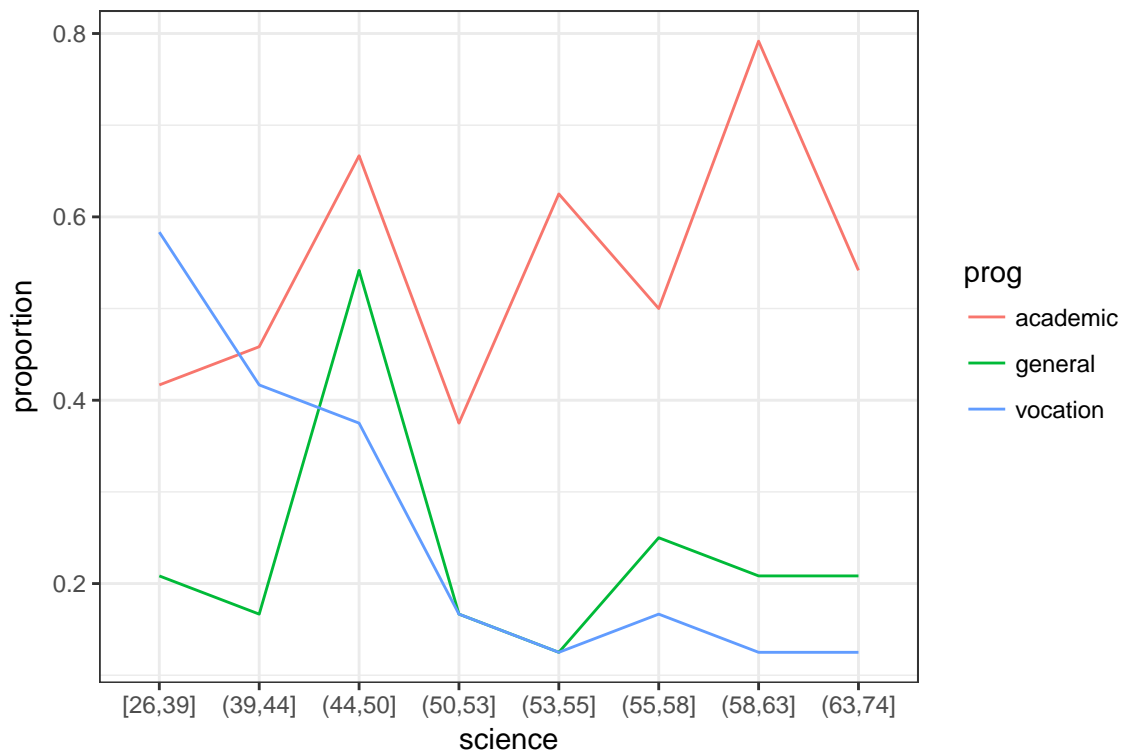
AIC: 347.6377

We can see that while students who have higher math and social studies scores tend to go to academic programs, the opposite is the case for science scores. This is not consistent with what we see in the data:

```

hsb.df %>%
  dplyr::mutate(science = cut_number(science, 8)) %>%
  dplyr::group_by(prog, science) %>%
  dplyr::summarise(y = n()) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(proportion = y / n()) %>%
  ggplot() +
  geom_line(aes(x = science, y = proportion, group = prog, colour = prog))

```



```
summary(multinom(prog ~ science, data = hsb.df))
```

```
# weights:  9 (4 variable)
initial value 219.722458
final value 196.328070
converged
```

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

```
Coefficients:
```

```
      (Intercept)      science
general -0.04375373 -0.01512511
vocation  2.83772488 -0.07082091
```

```
Std. Errors:
```

```
      (Intercept)      science
general  1.0102263 0.01880946
vocation  0.9561635 0.01895089
```

```
Residual Deviance: 392.6561
```

```
AIC: 400.6561
```

However, there is some correlation between math and science scores, which is probably the culprit. $\hat{\rho}_{math, science} \approx 0.631$. This is also probably why reading scores and socioeconomic status were omitted.

Interpretations:

- The probability of being in a general program when being in a private school and when subject scores are 0 is $\frac{e^{3.854}}{1 + e^{3.854} + e^{7.023}} \approx 0.0403$.
- The probability of being in a vocational program when being in a public school and when all subject

scores are 100 is $\frac{e^{7.023+1.788-13.70+4.214-8.672}}{1+e^{7.023+1.788-13.70+4.214-8.672}+e^{3.854+.6736-12.06+7.441-5.144}} \approx 8.6757 \times 10^{-5}$