Predicting Community Opinion on School Closures

Nick Koprowicz

May 06, 2017

Introduction

We set out to analyze data from Regression with Graphics by Hamilton, L.C., (1992). The data contained information on 153 individuals from a town in Vermont in six variables, described below:

Variable	Measure
gender	male or female.
lived	years respondent lived in the town.
kids	yes or no.
educ	respondents education (in years)
hsc	whether respondent attended meetings of the Health and Safety Committee, a citizens group organized in response to the towns contamination crisis.
contam	whether respondent believed his or her own property or water had been affected by chemical contamination
school	whether respondent thought that two schools with grounds contaminated by toxic waste should be closed until proven safe: open or closed.

Our goal is to explore the relationships between opinion on school closing and the other variables. To do this, we begin by looking at each variable individually.

Data Exploration

In order to get an idea of which of the variables has a relationship to opinion on school closing, we provide a series of tables and logistic regression models and interpret the results.

Variable 1. lived

The following table shows the mean number of years lived in the town within each opinion on closing the school.

Table 1. Lived vs. School

	Close	Open
Lived (mean)	13.68182	23.50575

It appears that people who have lived in the town longer want to keep the school open. Fitting a multinomial model gives the following estimates and standard errors:

```
Coefficients:
                 Values  Std. Err.
(Intercept) -0.46051084 0.26257879
d$lived      0.04100645 0.01214126
```

The P-value calculated from the estimate and standard error is 0.0007316205, which is significant. It appears that the number of years an individual has lived in the town has an impact on the opinion on school closing. The estimate 0.04100645 is an approximation of the amount that the log odds of someone choosing to keep the school open increases for each year they have lived in the town, so that the odds of someone who has lived in the town for 10 years want to keep the school open is exp(-0.46051084 + 0.04100645*10) = 0.9507.

Variable 2. education

The following table gives the mean number of years of education vs. opinion on school closing.

Table 2. Education vs. School

	Close	Open
Education (mean)	12.84848	13.03448

It appears that those in favor of keeping the school open have only a slightly higher mean number of years of education. Coefficients and standard errors of a multinomial model using education as the covariate follow:

```
Coefficients:
                 Values Std. Err.
(Intercept) -0.13460128 0.8888857
d$educ       0.03174757 0.0675845
```

The P-value computed from the estimate and standard error is 0.6385363, which is not significant. It looks like the number of years of education does not have an effect on the opinion on closing the school when considered as the only covariate.

Variable 3. gender

The following table shows the individuals of each gender and their opinion on school closing.

Table 3. Gender vs. School

	Close	Open
Male	20	40
Female	46	47

It looks like the females are evenly split, but men prefer to keep the school open. Fitting a multinomial gives the following estimates and standard errors:

```
Coefficients:
                Values Std. Err.
(Intercept)  0.02150356 0.2074023
d$gendermale 0.67163776 0.3435340
```

The P-value computed from the estimate and standard error is 0.05057315, which is barely not significant at the 0.05 level. We can calculate the odds that a male chooses to keep the school open by computing $\exp(0.02150356 + 0.67163776) \approx 2$, as we expect from the table. The odds that a female chooses to keep the school open is estimated to be $\exp(0.02150356) = 1.0217$.

Variable 4. kids

The following table shows the individuals with and without kids and their opinion on school closing.

Table 4. Kids vs. School

	Close	Open
No	20	43
Yes	46	44

Respondents who have kids appear to be evenly split in regards to keeping the school open, but respondents who do not have kids want to keep the school open. Fitting a model gives the following estimates and standard errors:

```
Coefficients:
               Values Std. Err.
(Intercept)  0.7654624 0.2706578
d$kidsyes   -0.8098587 0.3431064
```

The P-value associated with this estimate is 0.018256629, so it appears being a parent has a significant effect on school closing opinion. This model predicts that the odds that someone with kids chooses to keep the school open is exp(0.7654624 - 0.8098587) = 0.9568, and the odds that someone without kids chooses to keep the school open is exp(0.7654624) = 2.15.

Variable 5. hsc

The following table shows the number of individuals who attended or did not attend the meeting of the Health and Safety Committee (HSC) and their opinion on closing the school.

Table 5. Hsc vs. School

	Close	Open
No	28	78
Yes	38	9

Respondents who attended the HSC appear to want to close the school and those who did not attend the HSC appear to want to keep the school open. We compute the following estimates and standard errors by fitting a multinomial model:

```
Coefficients:
                Values Std. Err.
(Intercept)  1.024509 0.2203064
d$hscyes    -2.464862 0.4312320
```

The P-value is 1.091506e-08, which is very significant. There appears to be a strong relationship between attendance at the HSC and opinion on school closing. These estimates suggest that the odds that someone who attended the HSC wants to close the school is 1/exp(1.024509 - 2.464862) = 1/0.2368 = 4.222, and the odds that someone who did not attend the HSC wants to keep the school open is exp(1.024509) = 2.7857.

Variable 6. contam

The following table shows the number of individuals who do or do not believe that their own property had been contaminated and their opinion on closing the school.

Table 6. Contam vs. School

	Close	Open
No	36	74
Yes	30	13

It looks like those who believed their own property had been contaminated want to close the school, and those who believed their property had not been contaminated want to keep the school open. We compute the following estimates and standard errors:

```
Coefficients:
                Values Std. Err.
(Intercept)  0.7205214 0.2032017
d$contamyes -1.5567852 0.3892917
```

The P-value is 0.0000636054, which shows it's likely that there's a very strong relationship between those who believe their own property had been contaminated and opinion on school closing. The odds that someone who believes that their own property had been contaminated wants to close the school is 1/exp(0.7205214 - 1.5567852) = 2.307, and the odds that someone who doesn't think their property was contaminated wants to keep the school open is exp(0.7205214) = 2.055.

It appears that all of the measured variables have a relationship to the opinion of school closing except education. Now that we have some idea of which variables have an effect on the opinion of school closing, we proceed by fitting a model with several variables.

Model Fitting

Variable selection

Before considering interactions, we try to narrow down the variables in the model. Fitting a model with all the covariates, we get the following estimates and standard errors:

```
Coefficients:
                  Values  Std. Err.
(Intercept)  -1.89245480 1.36495550
d$gendermale  0.73073104 0.44226255
d$lived       0.04245891 0.01616080
d$kidsyes    -0.07308800 0.45690588
d$educ        0.16818646 0.09051775
d$hscyes     -2.27529163 0.48963796
d$contamyes  -1.18934406 0.46438786
```

Comparing the estimates to their standard errors, it appears that all covariates are significant except for `kids`. It's interesting that `educ` appears to be significant and `kids` does not. We reduce the model based on the AIC and we obtain the following estimates and standard errors:

```
Coefficients:
                  Values  Std. Err.
(Intercept)  -1.95226348 1.31351520
d$gendermale  0.72860858 0.44224206
d$lived       0.04336737 0.01516406
d$educ        0.16841452 0.09047756
d$hscyes     -2.28792288 0.48363027
d$contamyes  -1.18585781 0.4641462
```

The only variable that was removed from the model was `kids,` as we anticipated. The `educ` variable is kept in the model. As an alternative method of finding a reduced model, we start with an intercept and add variables one at a time, leaving the variable with the smallest P-value, as long as it is less than 0.25. After following this procedure, we arrive at the same model. It's interesting that `educ` is kept in both of these models even though there appeared to be no significant difference between mean years of education and school opinion. We remove `educ` from the current model and use the deviance to see if `educ` was significant in the fit of the model.

The residual deviance with `educ` in the model is 150.1884, and the deviance without `educ` is 146.5677. Since we removed one parameter from the model, the difference in deviance, $150.1884 - 146.5677 = 3.6207$ is distributed chi-square with 1 degree of freedom. The P-value is 0.05706486, which is not quite significant at the 0.05 level. On the side of caution, we keep `educ` in the model.

We also compute the deviance before and after removing `kids` from the model. The difference in the deviance is $146.5677 - 146.5421 = 0.0256$, distributed chi-square with 1 degree of freedom. The P-value is 0.8728811, which shows that we should keep `kids` out of the model for now. We will add `kids` as a covariate at a later time if we find that one of its interactions with another covariate is significant.

Interactions

We start with a model that contains all of the covariates except for `kids` and try fitting a model with interactions. One way to do this is to start with a model with all of the variables and their interactions, then reduce the model based on AIC. We will call this **Model 1**. We obtain the following estimates and standard errors:

Model 1.

```
Coefficients:
                               Values  Std. Err.
(Intercept)               -5.58104574 2.97588523
d$gendermale               3.89346219 1.53273332
d$lived                    0.03039897 0.01982181
```

```
d$educ                      0.47549911 0.22757344
d$hscyes                  -16.07050760 5.78212951
d$kidsyes                   6.81519470 3.25375159
d$contamyes                -2.58440468 0.98572945
d$gendermale:d$kidsyes     -3.94858849 1.64984190
d$educ:d$kidsyes           -0.52574526 0.25150545
d$kidsyes:d$contamyes       2.15573996 1.19985941
d$lived:d$hscyes            0.11366517 0.05454999
d$educ:d$hscyes             0.65830186 0.31675430
d$hscyes:d$kidsyes          3.02019664 1.98103629
```

We note that this model does contain the kids variable, which means that `kids` is significantly related to opinion on school closing through its interaction with other variables. In fact, the interaction between `kids` and every other covariate is included in this model. This prompts us to revisit the kids variable and how it's related to the other variables.

We find that even though individuals who have kids appear to be evenly split between opening and closing the school, 0.47058 of males with kids want to close the school compared with 0.3333 of males in general. Additionally, the proportion of individuals who do not believe their own property had been contaminated and wanted to close the school is 0.327, but the proportion of people who do not believe their property had been contaminated and have kids and want to close the school is 0.4375. The mean years of education among those without kids who want to close the school is 12 compared to 12.83721 for those who want to keep the school open in general. And finally, the proportion of individuals who attended the HSC and have kids and want to close the school is greater than the proportion of people who attended the HSC and want to close the school in general. Clearly it is important to leave kids in the model so that these interactions are part of the model.

However, this model is hardly parsimonious. We try another route of adding interactions. Rather than reducing the model based on AIC, we add interactions one at a time and keep the interaction with the lowest P-value, as long as it is less than 0.25. We continue this process any additional interaction has a P-value greater than 0.25. We call this **Model 2**. At the end of this process, we arrive at the following estimates and standard errors.

Model 2.

```
Coefficients:
                                Values  Std. Err.
(Intercept)                -5.58104574 2.97588523
d$gendermale                3.89346219 1.53273332
d$lived                     0.03039897 0.01982181
d$educ                      0.47549911 0.22757344
d$hscyes                  -16.07050760 5.78212951
d$kidsyes                   6.81519470 3.25375159
```

```
d$contamyes             -2.58440468 0.98572945
d$gendermale:d$kidsyes  -3.94858849 1.64984190
d$educ:d$kidsyes        -0.52574526 0.25150545
d$kidsyes:d$contamyes    2.15573996 1.19985941
d$lived:d$hscyes         0.11366517 0.05454999
d$educ:d$hscyes          0.65830186 0.31675430
d$hscyes:d$kidsyes       3.02019664 1.98103629
```

This model is similar to the one we arrived at by reducing based on AIC. Both models contain the interactions `gender*kids`, `educ*kids`, `kids*contam`, and `educ*hsc`. This is another valid model, though it also has many variables. The interactions seem to have a significant effect though, and should not be ignored. It is possible that in order to obtain a model that is more parsimonious, we could include the interactions that were common to both models. We will build a model with only these interactions and call it **Model 3**. The model has the following estimates and standard errors:

Model 3.

```
Coefficients:
                            Values Std. Err.
(Intercept)            -5.19359145 2.4136003
d$gendermale            2.96102366 1.0742102
d$lived                 0.05435087 0.0185018
d$educ                  0.39829853 0.1768856
d$hscyes               -7.41021271 3.3227836
d$kidsyes               6.08567463 2.8044620
d$contamyes            -2.59693481 0.9192492
d$gendermale:d$kidsyes -2.83996667 1.2038038
d$educ:d$kidsyes       -0.44813655 0.2080109
d$kidsyes:d$contamyes   2.38315980 1.1344534
d$educ:d$hscyes         0.35521229 0.2385909
```

Diagnostics and Evaluation

Confusion Matrices and Accuracy

In an effort to evaluate the effectiveness of our three models, we compute confusion matrices and the accuracy, based on fitting the entire data set.

Model 1.

```
          Reference
Prediction close open
     close    43   10
     open     23   77

               Accuracy : 0.7843
```

Model 2.

```
          Reference
Prediction close open
     close    44    6
     open     22   81

               Accuracy : 0.817
```

Model 3.

```
          Reference
Prediction close open
     close    43    8
     open     23   79

               Accuracy : 0.7974
```

We see that Model 2 has the best accuracy. It is interesting that Model 3 attained better accuracy than Model 1, even though it has less variables. Model 2 outperformed both models though, and looks to be optimal. In general, though, these models appear to have about equal accuracy.

We should use the likelihood ratio then to test if it makes sense to remove the variables from Model 1 and Model 2 that are not in common and arrive at model 3. Then, we should use the likelihood ratio test to see if we can remove any of the variables from model 3.

Likelihood Ratio Tests

We try removing the interactions from Model 1 that are not shared with Model 3 and perform a likelihood ratio test. We find that none of the variables are significant at the 0.05 level, but some are significant at the 0.10 level and all are significant at the 0.25 level. It's not clear that we can drop those interactions and consider only Model 3 moving forward.

Next, we try removing the interactions from Model 2 that are not shared with Model 3 and perform a likelihood ration test. Removing `lived*hsc` gives a P-value of 0.01468, which shows that that interaction is significant. Removing `kids*hsc` gives a P-value of 0.102, which isn't significant at the 0.10 level, but is less than 0.25. It's not clear that we can reject the larger Model 2 in favor of Model 3. We should continue to consider Model 1 and Model 2 moving forward.

We would then like to see if we can reduce Model 3 and end with something more parsimonious. We will remove variables one at a time and perform a likelihood ratio test, starting with the interactions. We find that `gender*kids, educ*kids,` and `kids*contam` are all significant at the 0.05 level. The P-value for `educ*hsc` is 0.1353, which is still less than 0.25. It's not clear that we can take any of the variables out of Model 3. Since these interactions involve each of the covariates, we must leave all of the covariates in the model too.

Pseudo-R^2

As a last effort to evaluate the fit of these models, we use McFadden's R^2. In linear regression, we use R^2 to see if our model fits the data well. McFadden's R^2 is a metric which can be used to evaluate logistic regression models. McFadden's R^2 is defined as $R^2 = 1 - \frac{log(L_c)}{log(L_{null})}$ where L_c is the maximized likelihood value from the fitted model, and L_{null} is the maximized likelihood value from the null model which contains only an intercept. We get the following values for each model:

Model 1: 0.4119506
Model 2: 0.4081483
Model 3: 0.3749830

This suggests that all these models fit about equally well. It's not surprising that the Model 1 fits best since it has the most covariates, followed by Model 2 and finally Model 3, which has the least number of covariates. These models seem to have about equal accuracy and fit about equally well. They are all plausible models that can be used to predict opinion on school closing well.

Conclusion

First, we examined each of the variables and their relationship to opinion on school closing individually.

We found that `hsc, lived,` and `contam` are very strongly related to the opinion on school closing. Individuals who attended the HSC meeting were more likely to want to close the school, and individuals who did not attend the HSC wanted to keep the school open. People who lived in the town longer were in favor of keeping the school open. And individuals who believed their own property was contaminated were in favor of closing the school, while people who did not believe their own property had been contaminated wanted to keep the school open.

The `gender` and `kids` variables were also related to the opinion on school closing, though less strongly. Males tended to be in favor of keeping the school open while females were evenly split. Individuals without kids were generally in favor of keeping the school open while individuals with kids were evenly split. Education did not appear to be related to the opinion on school closing.

When we fit a model with all of these covariates, we found that `kids` was not significant, but `educ` was. We used two methods of fitting the model with many covariates: reducing the model based on AIC, and adding covariates to the model one at a time based on P-value. We used the deviance to check that educ should be kept in the model and kids should not. Moving on to considering interactions, our base model contained all variables except `kids.`

We added interactions to the model in two ways: starting with all covariates and all possible interactions and reducing based on AIC, and starting with our base model and adding interactions one at a time based on P-value. We arrived at different models through these methods and called them **Model 1** and **Model 2**, respectively. We also considered a model with interactions in common to both of these methods and called it **Model 3**. These models are shown below.

Model 1.

```
Coefficients:
                            Values  Std. Err.
(Intercept)            -5.58104574 2.97588523
d$gendermale            3.89346219 1.53273332
d$lived                 0.03039897 0.01982181
d$educ                  0.47549911 0.22757344
d$hscyes              -16.07050760 5.78212951
d$kidsyes               6.81519470 3.25375159
d$contamyes            -2.58440468 0.98572945
d$gendermale:d$kidsyes -3.94858849 1.64984190
d$educ:d$kidsyes       -0.52574526 0.25150545
d$kidsyes:d$contamyes   2.15573996 1.19985941
d$lived:d$hscyes        0.11366517 0.05454999
d$educ:d$hscyes         0.65830186 0.31675430
d$hscyes:d$kidsyes      3.02019664 1.98103629
```

Model 2.

```
                            Values  Std. Err.
(Intercept)            -5.58104574 2.97588523
d$gendermale            3.89346219 1.53273332
d$lived                 0.03039897 0.01982181
d$educ                  0.47549911 0.22757344
d$hscyes              -16.07050760 5.78212951
d$kidsyes               6.81519470 3.25375159
d$contamyes            -2.58440468 0.98572945
d$gendermale:d$kidsyes -3.94858849 1.64984190
d$educ:d$kidsyes       -0.52574526 0.25150545
d$kidsyes:d$contamyes   2.15573996 1.19985941
d$lived:d$hscyes        0.11366517 0.05454999
d$educ:d$hscyes         0.65830186 0.31675430
d$hscyes:d$kidsyes      3.02019664 1.98103629
```

Model 3.

```
Coefficients:
                            Values Std. Err.
(Intercept)            -5.19359145 2.4136003
d$gendermale            2.96102366 1.0742102
d$lived                 0.05435087 0.0185018
d$educ                  0.39829853 0.1768856
d$hscyes               -7.41021271 3.3227836
d$kidsyes               6.08567463 2.8044620
d$contamyes            -2.59693481 0.9192492
d$gendermale:d$kidsyes -2.83996667 1.2038038
d$educ:d$kidsyes       -0.44813655 0.2080109
d$kidsyes:d$contamyes   2.38315980 1.1344534
d$educ:d$hscyes         0.35521229 0.2385909
```

We found that the there were significant interactions between `kids` and other covariates in all these models. We reexamined the data to see that there are interactions between `kids` and other variables and we interpreted these interactions.

Finally, for each of these three models, we performed diagnostics using confusion matrices to check accuracy, likelihood ratio tests to see if the models could be reduced, and McFadden's R^2 to check model fit. We found that regarding fit and accuracy, each model was about the same, with an accuracy of about 0.80. Model 2 slightly outperformed the others in accuracy, while Model 1 had the best fit. The benefit of Model 3 is that it is more parsimonious than the other models. Finally, likelihood ratio tests showed that no variables could be safely removed from these models.

It is worth noting the assumptions for logistic regression in this situation are met; each independent variable can take only one value, and the outcome is binary. There may be a limitation in considering `lived` and `educ` as continuous variables, since they only take discrete values. Our sample size is too small, however, to add them to a model as categorical variables. It may be worth exploring this option further; we may assign a few discrete values for years of education in certain ranges, and then include `educ` in the model as categorical, for example. Some primary experimenting with this process did not appear to yield better results. Further limitations in these models may exist as a result of our relatively small sample size.

We finish our report by providing some examples of how to use these models to predict opinion on school closing.

Examples of Implementation:

Example 1.

We use each model to estimate the odds that a female who has lived in the town for 15 years and has kids wants to keep the school open.

For Model 1, we compute the odds as exp(-5.58105 + 0.03039897(15) + 6.81519470) = 5.420181, a virtual certainty. Model 2 gives exactly the same value, since the estimates for those coefficients is the same. Using Model 3, we get exp(-5.19359145 + 0.05435087*(15) + 6.08567463) = 5.514308.

Example 2.

We use each model to estimate the odds that a male who has 12 years of education and believes their property has been contaminated wants to keep the school open.

Using Model 1, we compute exp(-5.58104574 + 3.89346219 + 0.47549911*12 - 2.58440468) = 4.195452. Again, Model 2 gives exactly the same value. For Model 3, we get exp(-5.19359 + 2.96102366 + 0.39829853*12 - 2.59693481) = 0.9513067. So Model 1 and Model 2 predict that an individual with these characteristics would be more likely to want to keep the school open, while Model 3 predicts that they would be more likely to want to keep the school closed.

Example 3.

We use Model 3 to predict the odds that a male who has lived in the town for 6 years, has 13 years of education, has kids, and attended the HSC meeting wants to keep the school open.

We compute the odds as exp(-5.19359145 + 2.96102366 + 0.05435087*6 + 0.39829853*13 + 6.08567463 - 7.41021271 - 2.83996667 - 0.44813655*13 + 0.35521229*13) = 0.1223216. So it is almost certain that someone with these characteristics would want to close the school.

Example 4.

We use Model 1 to predict the odds that someone with kids who attended the HSC and believes their property has been contaminated wants to keep the school open.

We compute exp(-5.58104574 + 6.81519470 - 2.58440468 - 16.07050760 + 2.15573996 + 3.02019664) = 4.809936e-06. There is essentially a probability of 0 that someone with these characteristics would want to keep the school open.

Bibliography

[1] Hamilton, L.C., (1992). Regression with Graphics. Belmont CA: Duxbury.

Appendix

```
#########################################################################
#*************************** CODE FOR FINAL PROJECT*********************#
#########################################################################

# Load the data

d = read.table("http://inside.mines.edu/~wnavidi/math436536/projects/Hamilton.txt", header = TRUE)
attach(d)

##################################
# Let's just investigate the data!
##################################


# Check for association between years lived in town and open/close

tapply(d$lived, d$school, mean)

# It looks like people who have lived in the town longet want to keep the school open



# Check for association between education and open/close

tapply(d$educ, d$school, mean)

# There doesn't appear to be a difference


# Check for an association between gender and open/close

table(d$gender, d$school)

# It looks like the females are evenly split, but the men preferred
# to keep the school open



# Check for an association between kids and school

table(d$kids, d$school)

# People who have kids are evenly split, but people who don't have kids prefer
# to keep the school open



# Check for an association between hsc and school

table(d$hsc, d$school)

# People who went to the hsc want to close the school, but those who didn't prefer to
# keep the school open



# check for an association between contam and school

table(d$contam, d$school)

# If people believed that their own property had been contaminated, they preferred
# to close, but if people didn't believe their property had been contaminated, they
# wanted to keep it open.



###############################################
# Let's fit a multinomial model with all the variables:
# Note that although 'lived' and 'educ' are discrete, we cannot consider
# them as factor variables because we do not have enough data to fit that many coefficients

library(nnet)
out = multinom(d$school~d$gender + d$lived + d$kids + d$educ + d$hsc + d$contam)
summary(out)

###############################################
# Let's use AIC to eliminate variables from the model
outa = step(out)

# kids is the only variable that's eliminated

# Let's look at the fit of the minimum AIC model:
out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam)
summary(out)

###############################################
# Let's compute P-values for the coefficients:
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# all the coefficients seem to be significant, even with the full model, although
# kids does have the highest P-value
```

```
#*******************************************************************************************#
#*******************************************************************************************#

# Let’s fit each variable, one at a time, and compute P-values

out = multinom(d$school ~ d$gender)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# gender appears to be significant

out = multinom(d$school ~ d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# years lived is very significant

out = multinom(d$school ~ d$kids)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# kids seems to be significant

out = multinom(d$school ~ d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# education DOES NOT appear to be significant

out = multinom(d$school ~ d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# hsc is very significant

out = multinom(d$school ~ d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

# contam appears to be very significant

#*******************************************************************************************#
#*******************************************************************************************#

# It’s interesting that kids was kicked out of the model when using AIC and
# education was kept in, although this is not the case when using P-values.
# This is probably something that can be explored with the deviance

# Let’s see what happens when we start with the model without kids and then
# see what happens to the deviance when we put kids back in the model

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam)
summary(out)

# The residual deviance is 146.5677

# Now, let’s look at the model with kids back in it

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$kids)
summary(out)

# The residual deviance is 146.5421

# The decrease in deviance is: 0.0256 with 1 degree of freedom
146.5677 - 146.5421

# The P-value is:
1 - pchisq(0.0256, 1)

# So there is no justification for adding kids back into the model.

# Now, let’s consider the model without kids or education and see what happens
# when we put education back in the model

out = multinom(d$school~d$gender + d$lived + d$hsc + d$contam)
summary(out)

# The residual deviance is 150.1884

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam)
summary(out)

# The residual deviance is 146.5677

# The decrease in the deviance is: 3.6207
150.1884 - 146.5677

# The P-value is:
```

```
1 - pchisq(3.6207, 1)

# The P-value is just greater than 0.05. We'll keep education in the model.

# (Note that when we performed this test, we had already removed kids from the
# model, but even if we had left kids in the model, the P-value is not very different.)

out = multinom(d$school~d$gender + d$lived + d$hsc + d$contam + d$kids)
summary(out)

# The residual deviance is 150.147

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$kids)
summary(out)

# The residual deviance is 146.5421

# The decrease in the deviance is: 3.6207
150.147 - 146.5421

# The P-value is:
1 - pchisq(3.6049, 1)

# The P-value is still just greater than 0.05. We'll keep education in the model.

#********************************************************************************************#
#********************************************************************************************#

# It's probably a good idea to ask Navidi about why the P-value for education
# was so high, when it really looks like it should be kept in the model

# So it seems like the full model has everything except kids


# Is education correlated with something else?

plot(d$lived, d$educ) # doesn't appear to be a relationship

tapply(d$educ, d$gender, mean) # doesn't appear to be a relationship

tapply(d$educ, d$kids, mean) # doesn't appear to be a relationship

tapply(d$educ, d$hsc, mean) # doesn't appear to be a relationship

tapply(d$educ, d$contam, mean) # doesn't appear to be a relationship

#********************************************************************************************#
#********************************************************************************************#


# Let's try fitting a model by adding variables one at a time

out = multinom(d$school ~ d$gender)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# gender appears to be significant, let's add another vriable and see what happens

out = multinom(d$school ~ d$gender + d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# the p-value for gender went up, but still below 0.25.  The P-value for lived is very low

out = multinom(d$school ~ d$gender + d$lived + d$kids)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# The P-value for kids is high, greater than 0.25, let's remove it from the model and add something else

out = multinom(d$school ~ d$gender + d$lived + d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# All P-values are less than 0.25

out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# All P-values get very small when hsc is addes to the model -> Apparently there is a strong association with HSC

out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# All P-values are significant.  Education should be kept in the model

# Now, try interactions:
out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived)
```

```
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# Keep gender*lived

out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# don’t add gender*educ
out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + gender*educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# don’t add it

out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# don’t add it

out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

# don’t add it

out = multinom(d$school ~ d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$lived*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2 # P = 0.05
p

#*******************************************************************************************#
#*******************************************************************************************#

# Now let’s try adding interactions to the model

install.packages(’caret’)
require(caret)

library(nnet)
out = multinom(d$school~d$gender + d$lived + d$kids + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$kids + d$gender*d$educ + d$gender*d$hsc+
                 d$gender*d$contam + d$lived*d$kids + d$lived*d$educ + d$lived*d$hsc + d$lived*d$contam + d$kids*d$educ + d$kids*d$hsc + d$kids*d$contam +
                 d$educ*d$hsc + d$educ*d$contam + d$hsc*d$contam)
summary(out)

###############################################
# Let’s use AIC to eliminate variables from the model
outa = step(out)


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$educ + d$gender*d$hsc+
                 d$gender*d$contam + d$lived*d$educ + d$lived*d$hsc + d$lived*d$contam + d$kids*d$hsc +
                 d$educ*d$hsc + d$educ*d$contam + d$hsc*d$contam)

outa = step(out)

summary(outa)


###############################################
# Let’s try adding interactions one at a time


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 4.577331e-02


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 5.696440e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
#  0.7199851814

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$contam)
```

```
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 9.351619e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 2.588432e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$kids*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.3522227366

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$kids*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 8.265897e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$kids*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.736400109

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$kids*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 9.946393e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$lived*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 1.293124e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$lived*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.070647901


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$lived*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 5.148044e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$educ*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.321255320

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$educ*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 8.996138e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$hsc*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 6.385609e-01


################## The Interaction with the smallest P-value is gender*kids.  Let's add that one and try the others.

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$gender*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 1.375130e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$gender*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$gender*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$gender*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
```

```
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 4.567487e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 6.602128e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$lived*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 8.044017e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$lived*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 6.131900e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$lived*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$educ*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$educ*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$contam*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

############################ The next best interaction is kids and education.  Let's add that one and try the others

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$gender*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 2.063672e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$gender*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$gender*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$gender*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$hsc)
```

```
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 6.479929e-02

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$lived*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$lived*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 1.304956e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$lived*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$educ*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$hsc*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

################################## The next best one is kids*contam.  Let's add that one and try the others


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam + d$gender*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$gender*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$gender*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$gender*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$kids*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$kids*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc )
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 2.028830e-01

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
```

```
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$educ*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.136542052

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ +d$kids*d$contam + d$hsc*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

##################### The next best one is lived*hsc Let's add that one and try the others


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  + d$gender*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$gender*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$gender*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$gender*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$kids*d$lived)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$kids*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.172632531

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$lived*d$educ)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$lived*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc  +d$educ*d$hsc)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.042077253

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ +d$kids*d$contam +d$lived*d$hsc  + d$hsc*d$contam)
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

####################################### The next best one is hsc*educ.  Let's add that one and go from there

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  + d$gender*d$
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$gender*d$e
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$gender*d$h
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
```

```
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$gender*d$c
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$liv
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# 0.127370398

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$lived*d$ed
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$lived*d$co
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ +d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  + d$hsc*d$cont
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

############# The only one that works is kids*hsc.  Let's add that one and go from there.




out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ +d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc   +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p
# no
```

```
################################## That's it! So our final model is:


out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc
z = summary(out)$coefficients/summary(out)$standard.errors
p = (1 - pnorm(abs(z), 0, 1))*2
p

################################## Let's try with the model with the interactions
# that were only in common to both models:

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam + d$educ*d$hsc)


########## To check:

probabilities = predict(out)

count = 0
for (i in 1:length(probabilities)){
  if(probabilities[i] == d$school[i]){count = count + 1}
}
count/length(probabilities)


# The percent of the add one in model is 0.8169935
# The percent of the AIC reduced model is 0.7973856
# The percent of the model with interactions in common to both 0.7973856
# The percent with no interactions at all is 0.7712418

#*******************************************************************************************#
#*******************************************************************************************#

#################################
# Performing diagnostics


# Model 1: Reduced by AIC

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$kids +
                 d$lived*d$educ + d$kids*d$educ + d$kids*d$contam + d$educ*d$contam + d$educ*d$hsc)

# Model 2: Add 1 in model

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$hsc

# Model 3: interactions that are in common to both models

out = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam + d$educ*d$hsc)

# The percent of the add one in model is 0.8169935
# The percent of the AIC reduced model is 0.7973856
# The percent of the model with interactions in common to both 0.7973856
# The percent with no interactions at all is 0.7712418

probabilities = predict(out)

################################## Confusion matrix

confusionMatrix(probabilities, d$school)

################################## Pseudo R^2

install.packages('pscl')
require(pscl)

pR2(out)

################################## likelihood ratio tests

# Model 1:

model1 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$kids +
                    d$lived*d$educ + d$kids*d$educ + d$kids*d$contam + d$educ*d$contam + d$educ*d$hsc)

# remove gender*lived

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$kids +
                    d$lived*d$educ + d$kids*d$educ + d$kids*d$contam + d$educ*d$contam + d$educ*d$hsc)

lrtest(model1, model2) # P = 0.05107

# remove lived*educ

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$kids +
                    d$kids*d$educ + d$kids*d$contam + d$educ*d$contam + d$educ*d$hsc)


# remove educ*contam

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$contam + d$gender*d$lived + d$gender*d$kids +
                    d$lived*d$educ + d$kids*d$educ + d$kids*d$contam + d$educ*d$hsc)


lrtest(model1, model2) # P = 0.1454
```

```
# Model 2:

model1 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc  +d$kids*d$

# remove lived*hsc

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$educ*d$hsc  +d$kids*d$hsc)

lrtest(model1, model2) # P = 0.01468

# remove kids*hsc

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam +d$lived*d$hsc +d$educ*d$hsc)

lrtest(model1, model2) # P = 0.102


# Model 3:

model1 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam + d$educ*d$hsc)

# remove gender*kids

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$kids*d$educ + d$kids*d$contam + d$educ*d$hsc)

lrtest(model1, model2) # P = 0.01

# remove educ*kids

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$contam + d$educ*d$hsc)

lrtest(model1, model2) # P = 0.02198

# remove kids*contam

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$educ*d$hsc)

lrtest(model1, model2) # P = 0.02721

# remove educ*hsc

model2 = multinom(d$school~d$gender + d$lived + d$educ + d$hsc + d$kids + d$contam + d$gender*d$kids + d$kids*d$educ + d$kids*d$contam)

lrtest(model1, model2) # P = 0.1353

################################## K-fold cross validation ?????????

ctrl = trainControl(method = "repeatedcv", number = 10, savePredictions = TRUE)
out = train(school~gender + lived + educ + hsc + kids + contam + gender*kids + kids*educ + kids*contam + educ*hsc,
            data = d, method = "multinom", trControl = ctrl)

probabilities = predict(out, d)



################################## Exploring how kids is related to other variables

b = d[d$kids == ’yes’,]

mean(b$educ)
mean(d$educ)

mean(b$lived)
mean(d$lived)

dim(b[b$hsc == ’yes’,])[1]/dim(b)[1]
dim(d[d$hsc == ’yes’,])[1]/dim(d)[1]

dim(b[b$contam == ’yes’,])[1]/dim(b)[1]
dim(d[d$contam == ’yes’,])[1]/dim(d)[1]

dim(b[b$gender == ’male’,])[1]/dim(b)[1]
dim(d[d$gender == ’male’,])[1]/dim(d)[1]

# This is what you’re looking for!

####### kids*gender

b = d[d$kids == ’yes’ & d$gender == ’male’,]
c = d[d$gender == ’male’,]
dim(b[b$school == ’close’,])[1]/dim(b)[1]
dim(c[c$school == ’close’,])[1]/dim(c)[1]

b = d[d$kids == ’yes’ & d$gender == ’female’,]
c = d[d$gender == ’female’,]
dim(b[b$school == ’close’,])[1]/dim(b)[1]
dim(c[c$school == ’close’,])[1]/dim(c)[1]

######## kids*contam

b = d[d$kids == ’yes’ & d$contam == ’yes’,]
c = d[d$contam == ’yes’,]
dim(b[b$school == ’close’,])[1]/dim(b)[1]
dim(c[c$school == ’close’,])[1]/dim(c)[1]
```

```
b = d[d$kids == 'yes' & d$contam == 'no',]
c = d[d$contam == 'no',]
dim(b[b$school == 'close',])[1]/dim(b)[1]
dim(c[c$school == 'close',])[1]/dim(c)[1]


####### kids*hsc

b = d[d$kids == 'yes' & d$hsc == 'yes',]
c = d[d$hsc == 'yes',]
dim(b[b$school == 'close',])[1]/dim(b)[1]
dim(c[c$school == 'close',])[1]/dim(c)[1]

b = d[d$kids == 'no' & d$hsc == 'yes',]
c = d[d$hsc == 'yes',]
dim(b[b$school == 'close',])[1]/dim(b)[1]
dim(c[c$school == 'close',])[1]/dim(c)[1]

b = d[d$kids == 'yes' & d$hsc == 'no',]
c = d[d$hsc == 'no',]
dim(b[b$school == 'close',])[1]/dim(b)[1]
dim(c[c$school == 'close',])[1]/dim(c)[1]

####### kids*educ

b = d[d$kids == 'yes',]
tapply(b$educ, b$school, mean)
tapply(d$educ, d$school, mean)

b = d[d$kids == 'no',]
tapply(b$educ, b$school, mean)
tapply(d$educ, d$school, mean)
```