Generalized Linear Models.

Question 4

A title :

Executive Summary

I conducted an observational study to examine factors thought to contribute to student absence from school in Largetown, Australia. Data were collected on student absence over the course of one school year for a sample of students. In addition to the number of school days missed, the grade level, ethnicity, gender, and speed of learning were collected because these factors are thought to influence absence from school.

Upon examination of the data, I determined that in addition to any of the factors suspected of influencing the number of school days missed, there were individual differences between students. I thus used a model that included such individual effects.

I found that, in addition to the individual student effects, both grade level and ethnicity explained some of the variation in the number of days missed. However, the effects of grade level and ethnicity were interdependent, meaning that the effect of grade level depended on the ethnicity of the student, and *vice-versa*. Gender and learning speed were not found to have any significant association with student absence, at least on average.

This report includes predictions of the number of days missed by ethnicity and grade level, as well as comparisons of the rate of absence between ethnicity-grade level groups.

Title :

1 Introduction

Data were collected on student absences in Largetown, Australia. In addition to the number of days of class missed, the grade level, ethnicity, gender, and learning speed were also recorded. Each of these factors are suspected of influencing the number of school days missed. I was interested the association between these factors and the rate of student absence from school. These associations might be independent of other factors, or they may depend on the level of other factors.

We will begin by assuming that the observed number of days missed is a random process based on some underlying rate of absence. Under such an assumption, the distribution of the number of days missed will follow a Poisson distribution, with parameter λ equal to the underlying rate of absence. Further, I will model this parameter using a generalized linear model of the form

$$\log(\lambda_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{1}$$

for student i, with factor combination $\boldsymbol{x}_i$. Then, given the rate for student i (i.e. λ_i), the number of days missed Y_i follows

$$Y_i \sim \text{Poisson}(\lambda_i) \tag{2}$$

We will be looking for the best combination of explanatory factors $\boldsymbol{x}_i$ to predict the number of days missed for every student in the dataset. We will then check to make sure that our assumed model is consistent with the data.

2 Preliminary analysis

I began by examining the data graphically, beginning with the marginal distribution of days absent for each factor (figure 1). The number of days missed was examined on the log-scale due to the assumed model form (equation 1). Although there are some differences in the number of days missed for each of the factors, there is a lot of variability within each group. For example, although aboriginal students miss on average twice as many days of school as non-aboriginals, each group covers nearly the same number of days missed (from zero to more than $2^6 = 64$).

The large amount of variation within groups could be explained by interactions between the factors. If this were the case, then we should observe markedly less variation when we divide the students into as many groups as possible based on the levels of the four factors (figure 2). Examining the data in this way, we notice at most a slight decrease in variation within groups. We also notice that there are some groups with few or no students - in particular, there are no average learners at the F3 grade level! This suggests a bias in the collection of data, but we won't address this bias here due to a lack of available information on data collection procedures.

Suspecting variation beyond what is expected from the model in equations 1 and 2, I wanted to examine the relationship between the means and variances

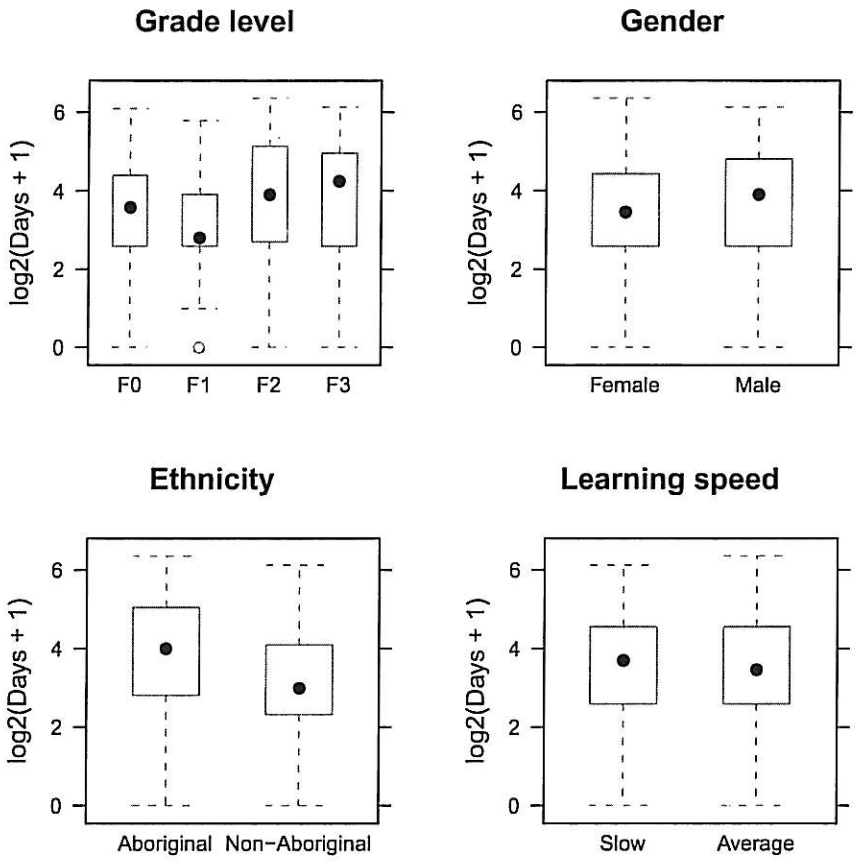

Figure 1: Marginal distribution for the number of school days missed, plotted on the log base 2 scale for ease of interpretation. The points denote the mean of the data, the boxes cover the interquartile range, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range.

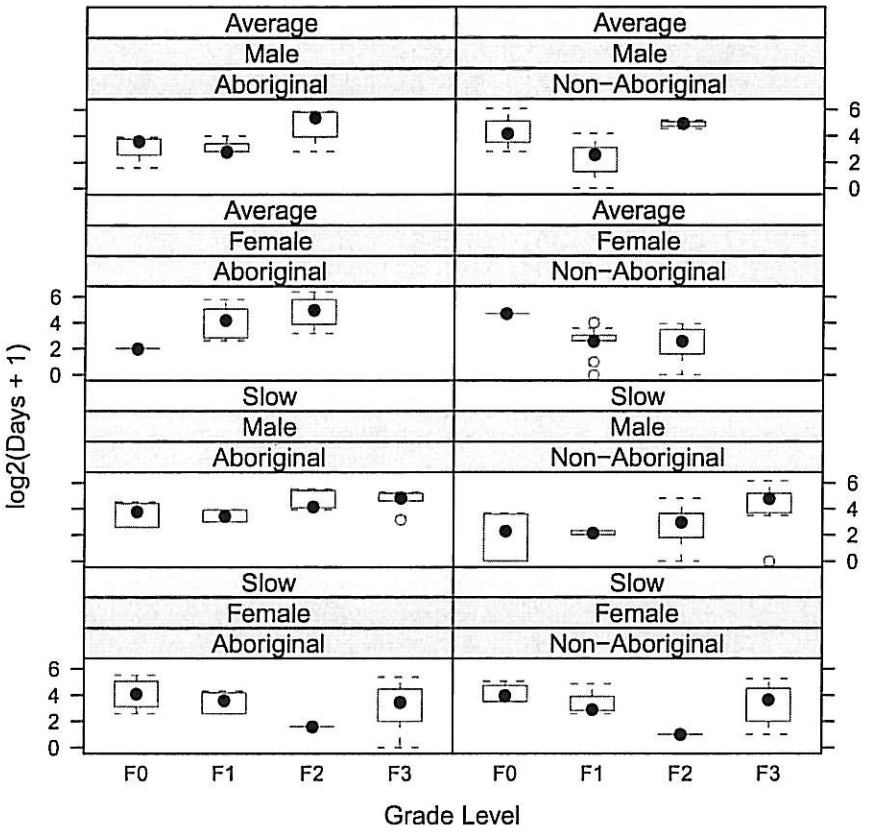

Figure 2: Distributions of the number of school days missed, plotted on the log base 2 scale for ease of interpretation. The points denote the mean of the data, the boxes cover the interquartile range, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range.

of the groups. Under the model assumed, the means and variances should be equal. I examined the means and variances within for each group from figure 2 (figure 3). There is obviously way more variation than expected within each group - a violation of the assumed model. It is clear that we should modify our model to allow for extra student-level variation within groups.

A formal test of this extra variation is to examine the deviance from a fitted model of the type described by equations 1 and 2. The deviance should follow a χ^2 distribution with $n-p$ degrees of freedom, where n is the number of students and p is the number of parameters in the model (i.e. the length of β). Fitting the model with the most groups possible, I obtained a deviance of 1174 on 118 degrees of freedom. The p-value for this deviance statistic is essentially zero, indicating that it is extremely unlikely that the data are due to a process like the Poisson regression described by equations 1 and 2. Rather, we will have to modify our model to include this *over-dispersion* - that is, we need to include student-level variation in addition to group-level variation.

3 Models with over-dispersion

A popular approach for handling over-dispersion is to incorporate individual-level effects in addition to group level effects. This combination of effects on different experimental units is often referred to as a *mixed effects* model. In our case, the *random effects* describe the student-level variation, and the *fixed effects* the group-level variation.

The approach is to modify our model from equations 1 and 2. Equation 1 becomes

$$\log(\lambda_i) = \log(u_i) + \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{3}$$

where u_i are the individual student-level effects assumed to follow

$$u_i \sim \text{Gamma}\left(\frac{1}{\phi}, \phi\right) \tag{4}$$

i.e. the mean of the u_i is 1 and the variance is an unknown parameter ϕ. It is (somewhat) clear why the u_i are called random effects: they are thought to have arisen from a random process described by equation 4, and thus would take on different values were the experiment repeated. This is in contrast to the group-level fixed effects β, which are thought to be set by the experimenter, and would not differ were the experiment repeated.

The choice of a Gamma density in equation 4 is to simplify the likelihood, which turns out to be

$$L(y_i|\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} \frac{\Gamma(1/\phi + y_i)}{\Gamma(1/\phi) y_i!} \left(\frac{e^{x_i^T \beta} \phi}{1 + e^{x_i^T \beta} \phi}\right)^{y_i} \left(\frac{1}{1 + e^{x_i^T \beta} \phi}\right)^{1/\phi} \tag{5}$$

i.e. a negative binomial distribution where if $w_i = y_i + 1/\phi$ is the number of trials until $1/\phi$ successes, then

$$W_i \sim \text{NegBin}\left(\frac{1}{\phi}, \frac{1}{1 + \lambda_i \phi}\right)$$

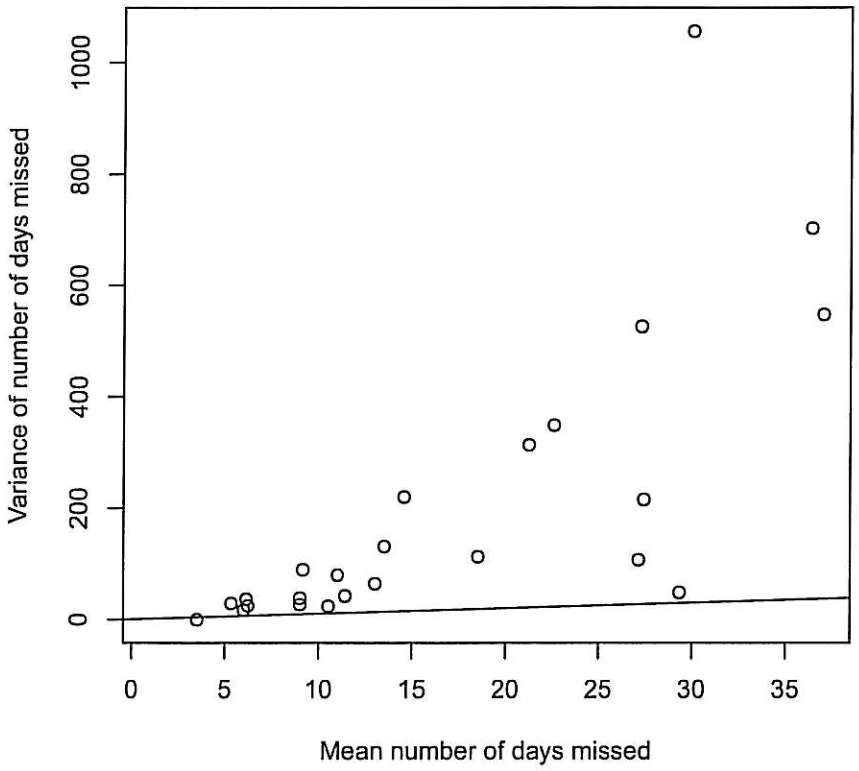

Figure 3: Means versus variances for the groups depicted in figure 2. The line is a one-to-one line, which is expected under the model assumed by equations 1 and 2.

Nice!

In any case, we can use this likelihood to estimate the model parameters β and ϕ from equation 5 without reference to any u_i. This will allow us to estimate the parameters of interest (the β) while accounting for the over-dispersion (the ϕ).

4 Model selection

In the absence of any information about which factors might affect school absence and in what way, my approach for selecting an appropriate model for the data was to follow an information theoretic approach by using Akaike's information criterion (AIC). The AIC is a measure of information entropy - we seek the model that gives the smallest AIC in order to minimize this entropy. Roughly speaking, it is a tradeoff between deviance and model complexity. The AIC is computed as

$$AIC = 2p - 2\log(L) \tag{6}$$

Notice that the AIC is itself a random quantity, and so we should consider models with similar AIC statistics to provide similar quality of fit.

I performed model selection as follows:

1. START with the simplest model (zero order: intercept only) as the best model.
2. Contruct candidate models by adding each of the effects at the next higher order, and removing each of the effects at the next lower order, preserving hierarchies. That is, for an interaction term to appear in a candidate model, all terms of lower order containing the interacting factors must also be included.
3. Compute the AIC statistic for each of the candidate models from step 2.
4. If any of the the candidate models has a smaller AIC than the previous best model, then choose the model with the smallest AIC statistic from among the candidates. Otherwise, STOP.
5. Repeat steps 2-4 until none of the candidates offer an improved AIC statistic.

In the results that follow, the *Age* factor refers to the grade level (F0, F1, F2, F3); the *Eth* factor refers to the ethnicity (Aboriginal, Non-aboriginal); the *Sex* factor refers to the gender (Female, Male); the *Lrn* factor refers to the learning speed (Slow, Average). Higher order terms are denoted by : for interactions, and * for complete crossings, e.g. Eth:Age denotes the interaction between Eth and Age, whereas Eth*Age denotes Eth + Age + Eth:Age.

The final model chosen was the one with ethnicity, grade level, and their interaction (table 1). I repeated the same model selection procedure, once starting with the most complex model instead of the empty model (table 2), another time

starting with an intermediate model that included all of the second order interactions (table 3), and a final time starting with an intermediate model that included all of the first order terms (table 4).

Table 1: Trace of the model selection procedure, beginning with the simplest (empty) model. The dispersion parameter ϕ for each model is also reported.

	Model	AIC	p	ϕ
1		1120	1	0.94
2	Eth	1111	2	0.86
3	Eth + Age	1106	5	0.80
4	Eth + Age + Eth:Age	1101	8	0.74

Table 2: Trace of the model selection procedure, beginning with the most complex (full) model. The dispersion parameter ϕ for each model is also reported.

	Model	AIC	p	ϕ
1	... + Eth:Sex:Lrn + Eth:Age:Lrn + Sex:Age:Lrn + Eth:Sex:Age:Lrn	1095	28	0.52
2	... + Eth:Sex:Age + Eth:Sex:Lrn + Eth:Age:Lrn + Sex:Age:Lrn	1093	26	0.52
3	... + Age:Lrn + Eth:Sex:Lrn + Eth:Age:Lrn + Sex:Age:Lrn	1089	23	0.54

Table 3: Trace of the model selection procedure, beginning with a model that included all second order interaction terms. The dispersion parameter ϕ for each model is also reported.

	Model	AIC	p	ϕ
1	... + Eth:Age + Eth:Lrn + Sex:Age + Sex:Lrn + Age:Lrn	1099	18	0.62
2	... + Eth:Lrn + Sex:Age + Sex:Lrn + Age:Lrn + Eth:Sex:Lrn	1091	19	0.58
3	... + Eth:Lrn + Sex:Age + Sex:Lrn + Age:Lrn + Eth:Sex:Lrn	1091	16	0.61

Given the results of these AIC calculations, we are left with a choice between several models that have very similar AIC values, but in some cases very different model complexities. Of those examined, the model with the smallest AIC statistic had p = 23 parameters: Eth + Sex + Age + Lrn + Eth:Sex + Eth:Age + Eth:Lrn + Sex:Age + Sex:Lrn + Age:Lrn + Eth:Sex:Lrn + Eth:Age:Lrn + Sex:Age:Lrn (table 2). By comparison, a model with p = 8 parameters: Eth + Age + Eth:Age had an AIC only 12 points (or about one percent) larger (table 1).

Notice that the AIC calculations omit the effect of the student-level variation: Comparing the deviance of these two models, the one with p = 23 has a deviance of 167 whereas the model with p = 8 has a deviance of 168. This suggests that

Table 4: Trace of the model selection procedure, beginning with a model that included all first order terms. The dispersion parameter ϕ for each model is also reported.

	Model	AIC	p	ϕ
1	Eth + Sex + Age + Lrn	1107	7	0.78
2	Eth + Sex + Age + Lrn + Sex:Age	1100	10	0.71
3	Eth + Sex + Age + Lrn + Sex:Age + Eth:Age	1095	13	0.65
4	Eth + Sex + Age + Lrn + Sex:Age + Eth:Age + Age:Lrn	1094	15	0.63

the validity of the additional complexity is spurious and might be just as well attributed to the individual-level random effects. We might in fact do a poorer job of predicting future data using the more complex model because we would be interpreting student-level variation as complex group-level variation when we ought not be doing so. I suggest we should prefer the simpler model with p = 8 parameters due to this possibility.

5 Parameter estimates and interpretation

I chose the best model from table 1 as the best combination of simplicity and fit to the data. A fit of this model to the data using the likelihood from equation 5 provided estimates in table 5. Missing from the table is the estimated student-level variance $\hat{\phi}$ =0.74.

Table 5: Parameter estimates for the fixed effects from the chosen model. Also reported are the standard errors, the value of the z-statistic assuming a normal distribution for the estimates, and the p-value associated with each test.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.6280	0.2495	10.54	0.0000
EthNon-Aboriginal	0.1311	0.3455	0.38	0.7044
AgeF1	0.1784	0.3195	0.56	0.5766
AgeF2	0.8267	0.3172	2.61	0.0092
AgeF3	0.3708	0.3337	1.11	0.2665
EthNon-Aboriginal:AgeF1	-0.9916	0.4394	-2.26	0.0240
EthNon-Aboriginal:AgeF2	-1.2392	0.4466	-2.78	0.0055
EthNon-Aboriginal:AgeF3	-0.1763	0.4636	-0.38	0.7038

In lay terms, the model chosen for the data says that there are differences in the number of school days missed when comparing students that have different ethnicities and that are at different grade levels, but that these differences are not independent from each other. In other words, the difference in the number of days of school missed between aboriginal and non-aboriginal students is different

by grade level, and the difference in the number of days of school missed between students in different grade levels is different depending on whether they are aboriginal or non-aboriginal.

The estimates for the fixed effects can be interpreted as average differences between groups, where groups are defined by those terms included in the model. This is because the average random effect within groups is defined to be zero. Using the estimates, I calculated the predicted number of days of school missed for each of the groups included in the model (table 6).

Table 6: Predicted number of school days missed as estimated from the chosen model. Also reported are the standard errors.

	Ethnicity	Grade level	Y_{pred}	$se(Y_{pred})$
1	Aboriginal	F0	13.85	3.45
2	Non-Aboriginal	F0	15.79	3.77
3	Aboriginal	F1	16.55	3.30
4	Aboriginal	F2	31.65	6.20
5	Aboriginal	F3	20.06	4.45
6	Non-Aboriginal	F1	7.00	1.29
7	Non-Aboriginal	F2	10.45	2.13
8	Non-Aboriginal	F3	19.18	4.13

Finally, for those interested in the relative rate of absence from school between the various groups in the model, I calculated the relative rates and 95 percent confidence intervals for comparisons between each of the groups. These are calculated by first estimating the log relative rate

$$\widehat{\log\left(\frac{\lambda_i}{\lambda_j}\right)} = x_i^T\hat{\beta} - x_j^T\hat{\beta} \tag{7}$$

for students from two groups indexed by i and j, $i \neq j$. The variance of the estimate is obtained from the estimated covariance matrix for the $\hat{\beta}$ (denoted $\hat{V}$) using

$$\widehat{\mathrm{Var}}\left[\log\left(\frac{\lambda_i}{\lambda_j}\right)\right] = c^T\hat{V}c \tag{8}$$

where $c = x_i - x_j$. This variance was then used to construct 95 percent confidence intervals for $\log(\lambda_i/\lambda_j)$, which were subsequently exponentiated to obtain the relative rates and confidence intervals. Table 7 contains the estimates relative rates, and tables 8 and 9 contain the lower and upper 95 percent confidence bounds, respectively.

6 Model checking

The residuals are well-behaved; deviance residuals are within two deviance units nineteen times out of twenty, and do not show a pattern when plotted against

Table 7: Relative rate of school absence for the groups in table 6. The relative rates are listed as column versus row - e.g. the entry in column 2, row 1 is the relative rate of those in group 2 versus those in group 1.

	1	2	3	4	5	6	7	8
1		1.14	1.20	2.29	1.45	0.51	0.75	1.38
2	0.88		1.05	2.00	1.27	0.44	0.66	1.21
3	0.84	0.95		1.91	1.21	0.42	0.63	1.16
4	0.44	0.50	0.52		0.63	0.22	0.33	0.61
5	0.69	0.79	0.82	1.58		0.35	0.52	0.96
6	1.98	2.26	2.36	4.52	2.87		1.49	2.74
7	1.32	1.51	1.58	3.03	1.92	0.67		1.84
8	0.72	0.82	0.86	1.65	1.05	0.37	0.54	

Table 8: Lower 95 percent confidence interval bound for the relative rate of school absence for the groups in table 6. The lower bounds are listed as column versus row - e.g. the entry in column 2, row 1 is the lower bound for the relative rate of those in group 2 versus those in group 1.

	1	2	3	4	5	6	7	8
1		1.14	1.20	2.29	1.45	0.51	0.75	1.38
2	0.88		1.05	2.00	1.27	0.44	0.66	1.21
3	0.84	0.95		1.91	1.21	0.42	0.63	1.16
4	0.44	0.50	0.52		0.63	0.22	0.33	0.61
5	0.69	0.79	0.82	1.58		0.35	0.52	0.96
6	1.98	2.26	2.36	4.52	2.87		1.49	2.74
7	1.32	1.51	1.58	3.03	1.92	0.67		1.84
8	0.72	0.82	0.86	1.65	1.05	0.37	0.54	

?

Wrong table?

the fitted values (figure 4). As expected, the deviance residuals also do not show a pattern when plotted against each of the factors in the dataset (figure 5). The model appears to describe the data well. In particular, the choice of a more complex model, as suggested by choosing the smallest AIC statistic, seems unfounded. Our simpler model does not appear to leave out important trends, nor does it appear to have any eggregious lack of fit for any particular student.

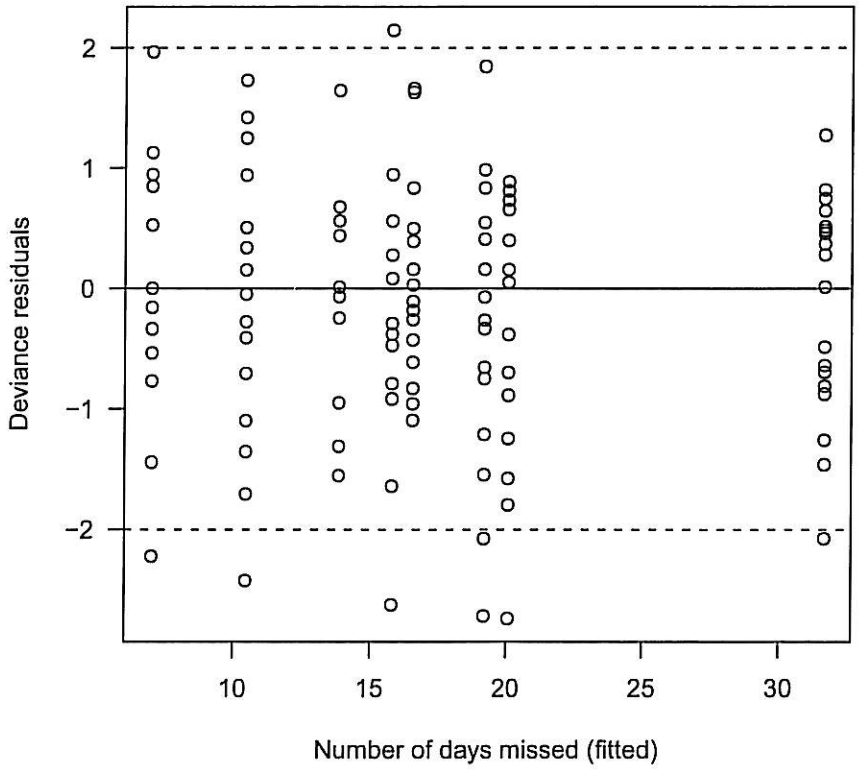

Figure 4: Deviance residuals plot. Residuals outside of the dashed lines are only expected to occur for one out of every twenty students assuming the model is correct. There are n = 146 students.

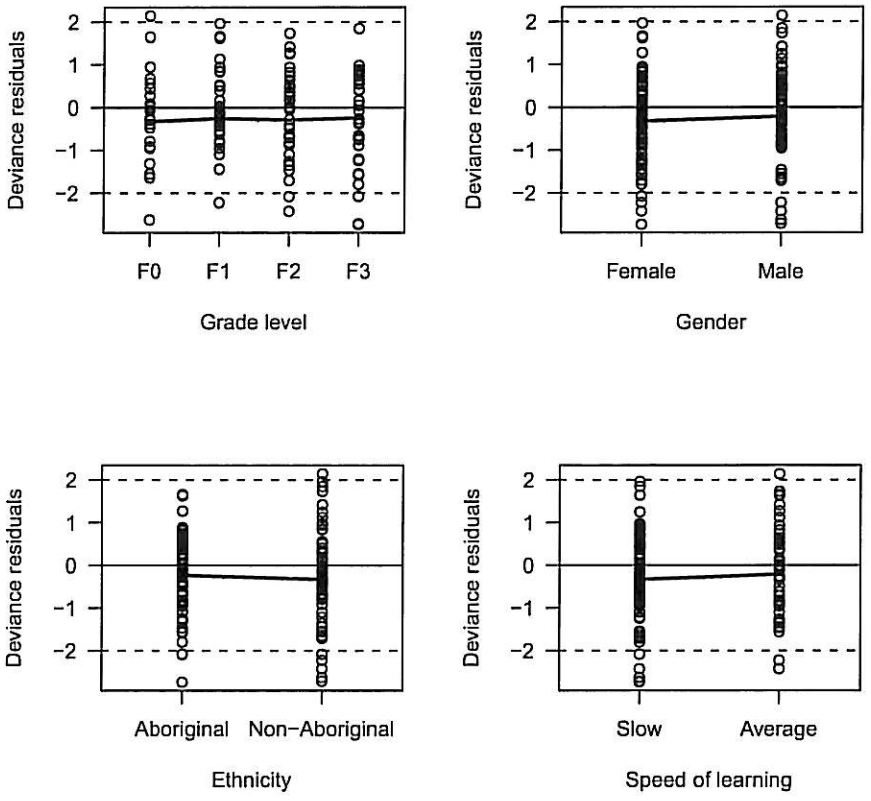

Figure 5: Deviance residuals versus each factor in the dataset. Residuals outside of the dashed lines are only expected to occur for one out of every twenty students assuming the model is correct. There are n = 146 students. The heavy line is a lowess smooth fit the points to detect trends.

Table 9: Upper 95 percent confidence interval bound for the relative rate of school absence for the groups in table 6. The upper bounds are listed as column versus row - e.g. the entry in column 2, row 1 is the upper bound for the relative rate of those in group 2 versus those in group 1.

	1	2	3	4	5	6	7	8
1		1.14	1.20	2.29	1.45	0.51	0.75	1.38
2	0.88		1.05	2.00	1.27	0.44	0.66	1.21
3	0.84	0.95		1.91	1.21	0.42	0.63	1.16
4	0.44	0.50	0.52		0.63	0.22	0.33	0.61
5	0.69	0.79	0.82	1.58		0.35	0.52	0.96
6	1.98	2.26	2.36	4.52	2.87		1.49	2.74
7	1.32	1.51	1.58	3.03	1.92	0.67		1.84
8	0.72	0.82	0.86	1.65	1.05	0.37	0.54	

7 Conclusion

I began by examining the data grouped at the finest scale, and noticed an amount of variation in the number of days of school missed far in excess of what is expected under the log-linear model described by equations 1 and 2. I therefore examined models that included student-level variation in the rate of school absence in addition to group-level variation. In terms of their ability to fit the observed data, relatively simple models for fixed effects had similar performance to more complex models. This suggested that much of the variance was due to student-level factors or other factors not considered. Thus, I chose a simpler model for the fixed effects to avoid overfitting and hopefully improve predictive performance.

Under the simpler model, I found that both grade level and ethnicity played a role in determining the average number of school days missed by students. However, the effects of grade level and ethnicity depended were not independent - rather, the effect of grade level depended on the ethnicity of the student, and the effect of ethnicity depeneded on the grade level. Gender and learning speed were not found to have any effect on student absence at the group level.

R code

This document was generated using Sweave and R. This section contains the R code chunks used throughout the document to generate statistics, graphics, and tables.

```
Writing to file final.R
```

```
###################################################
### chunk number 1:
###################################################
options(width=65)
rm(list=ls())
source("Sweave.R")
require(lattice)
school <- read.table( "school.txt", header = TRUE )
levels(school$Eth) <- c("Aboriginal","Non-Aboriginal")
levels(school$Sex) <- c("Female","Male")
levels(school$Lrn) <- c("Slow","Average")


###################################################
### chunk number 2:
###################################################
age.p <- bwplot( log2( Days + 1 ) ~ Age, data = school, main="Grade level" )
eth.p <- bwplot( log2( Days + 1 ) ~ Eth, data = school, main="Ethnicity" )
sex.p <- bwplot( log2( Days + 1 ) ~ Sex, data = school, main="Gender" )
```