Sta2201 Assignment 2

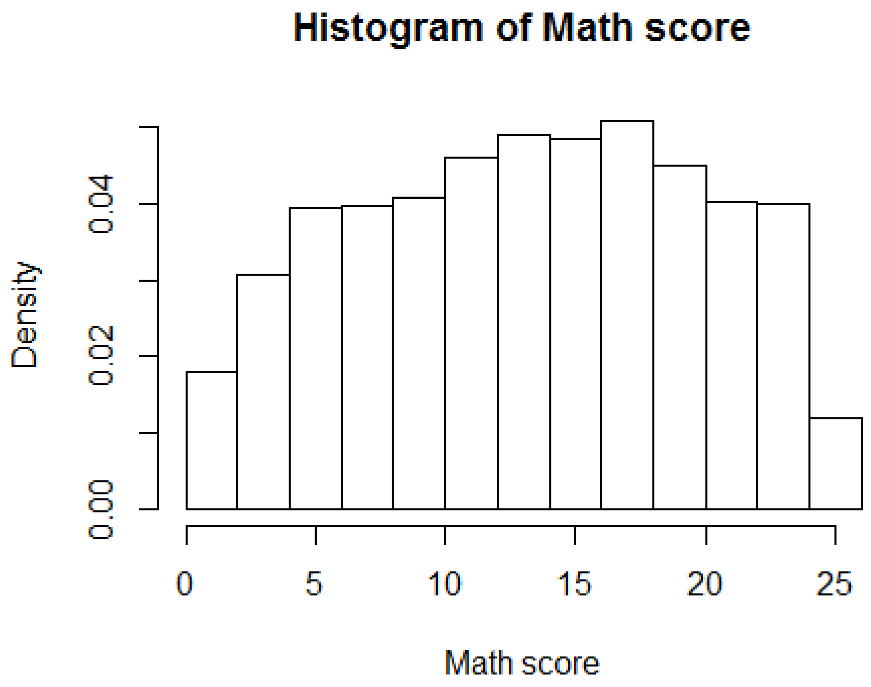
Jiahui Du

Feb 16, 2019

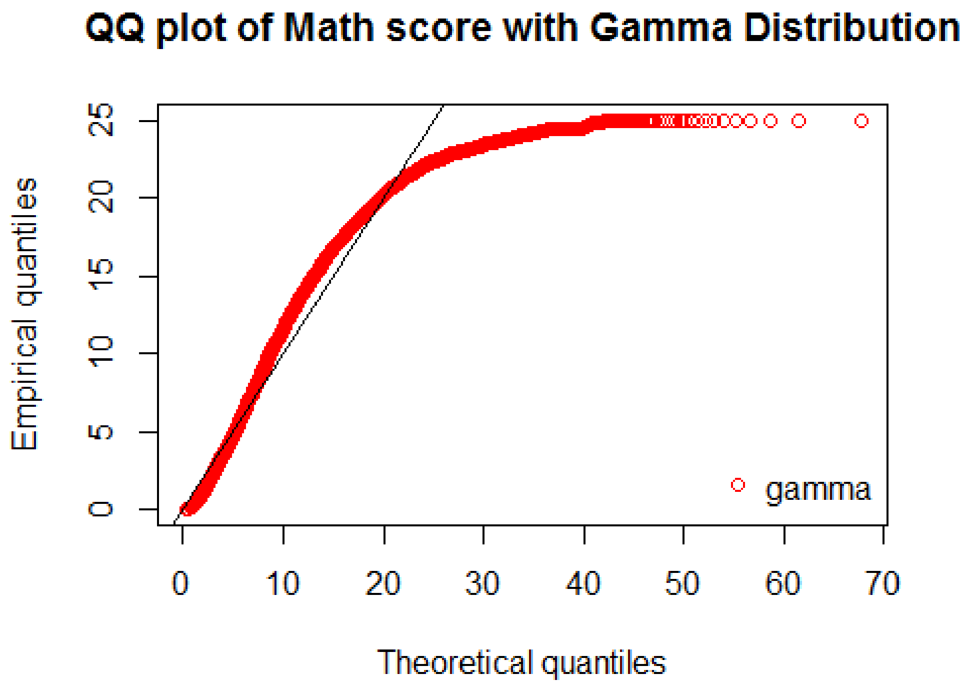
Student number: 998268556

Q1

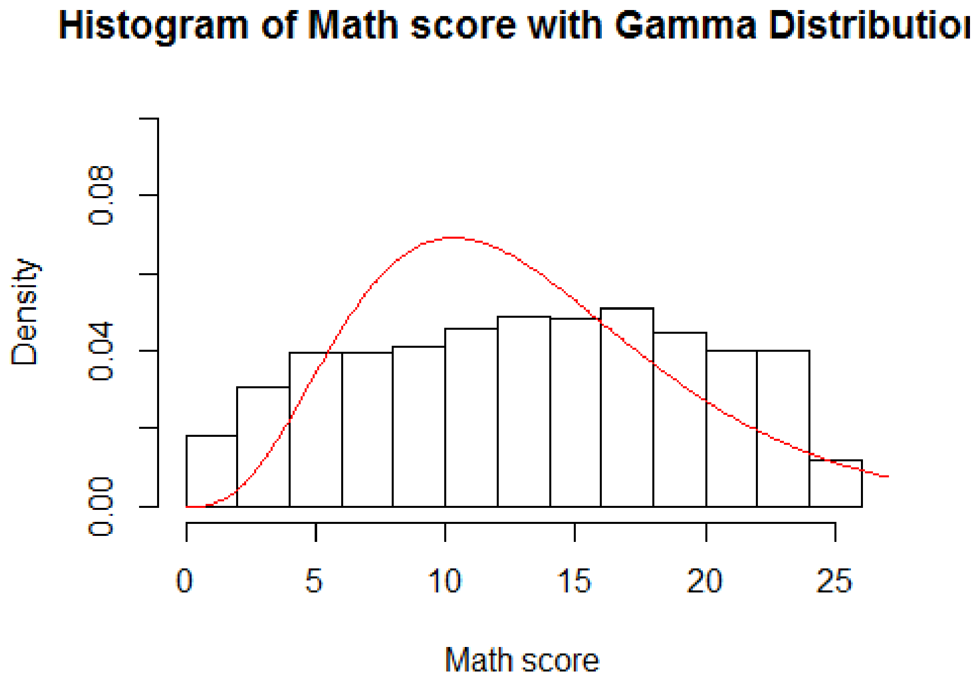
At first we will look at the shape of histogram of the data below in order to fit a proper model for Math score.



The Math score apears to be left skewed and it is always greater than zero. These evidenets may prevent us from using Normal distribution. And gamma may seem to fit the Math score with above mentioned features of the data.



QQ plot of Math score with Gamma Distribution



Histogram of Math score with Gamma Distributioi

After fitting the Gamma, a QQ-plot of gamma is shown above. We could see the model fit weill until when it reaches 25. It is because there is a maximum math score cut off at 25. From the histogram we could see the gamma model over weights around math score at 10.

In order to find out whether the differences between schools are greater than within-school variation, a model with school as random effect will be used. The model will give shape and scale values for the gamma thence we get the variance of test score for a typical school test score. Then we simulate the random intercept from a normal distribution with 0 mean and 0.103 standard deviation (from

VarCorr(model)) and use those simulated random intercept to simulate the test score. And we could find the sampled variance to estimate the variance of school average test score.

Variance of Random Effects

	2.5%	97.5%
Typical School	3.922	28.562
School Average	10.891	16.111

From the table above, we could see the whole 95% gamma quantile interval for school average is contained within the typical school's. Therefore we could conclude the within school variation is much bigger than the between school variation.

Q2

This is to study the effect of the F508 gene on the decline in lung function in individuals with cystic fibrosis over age. There are mainly two research hypotheses: the rate at which lung function declines for CF patients depends on the F508 gene and the effect of the F508 gene on lung function decline differs for females and males over age.

A model (Model 1) was suggested bt the medical scientist and it can be represented as:

Model 1

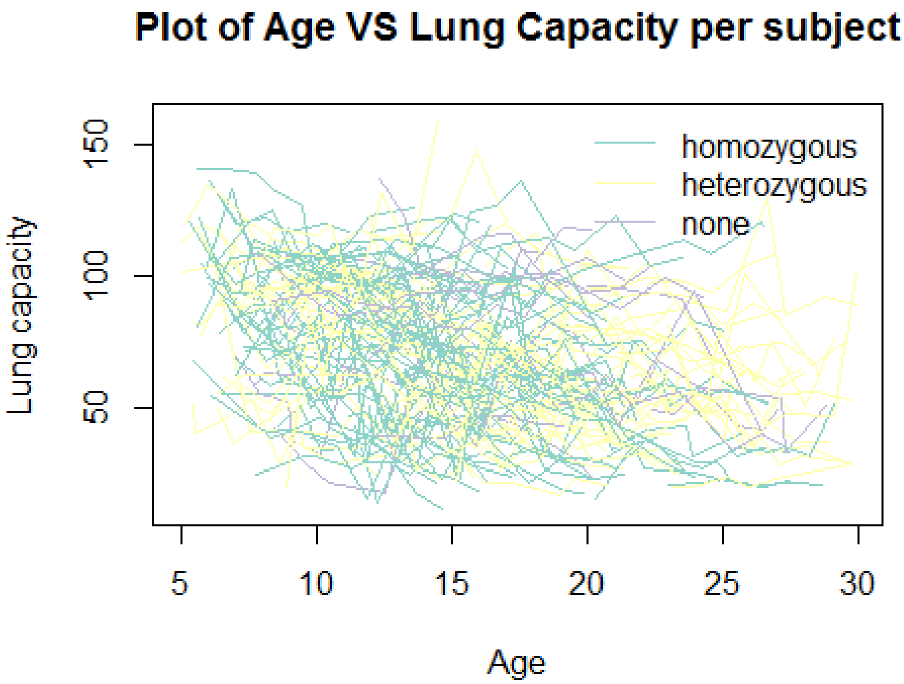
$$Y_{ij} \sim N(U_i + x_{ij}\beta, \tau^2)$$
$$U_i \sim N(0, \sigma^2)$$

where Ui is the random intercept for each subject ID. And the random intercept means that each subject may have different lung function. Also, the interaction term between Age, F508 and gender means the effect of one of the terms to lung function differs per different values of the other two terms. Similar idea applies to the two term interactions. And these interactions will help us answer the research hypotheses. Below are the results medical scientist created.

Model 1 - Random intercept

	Value	Std.Error	DF	t-value	p-value
Intercept	66.877	3.749	1306	17.840	0.000
Gender(Female)	-2.203	4.996	194	-0.441	0.660
F508(heterozygos)	6.269	5.001	194	1.254	0.212
F508(none)	7.157	7.348	194	0.974	0.331
Age	-1.754	0.251	1306	-6.989	0.000
Pseudoa(Yes)	-2.159	1.059	1306	-2.038	0.042
Gender(Female) * F508(heterozygos)	-6.275	7.031	194	-0.892	0.373
Gender(Female) * F508(none)	2.014	11.032	194	0.183	0.855
Gender(Female) * Age	0.071	0.352	1306	0.200	0.841
F508(heterozygos) * Age	0.744	0.344	1306	2.165	0.031

F508(none) * Age	1.610	0.534	1306	3.014	0.003
Gender(Female) * F508(heterozygos) * age	-0.998	0.495	1306	-2.018	0.044
Gender(Female) * F508(none) * age	-1.889	0.810	1306	-2.333	0.020



However, if we look at the plot above, the slope between age and lung capacity is not necessary to be the same given same gene and age for each subject. This may suggest a random slope in age or a correlation of age within same subject is needed. Therefore, there are two additional models that now we are considering. They are a model with random slope for age (model 2) and a model with serial correlation of lung capacity over age (model 3) and they are represented as below:

Model 2

$$Y_{ij} \sim N(U_{i1} + U_{i2}Age + x_{ij}\beta, \tau^2)$$
$$\begin{pmatrix} U_{i1} \\ U_{i2} \end{pmatrix} \sim MVN(0, \Gamma)$$

The random slop for age from Model 2 indicates that there is a different rate of change in lung capacity as age increases per subject. Below is the result.

Model 2 - Random Slope

	Value	Std.Error	t-value	p-value
Intercept	68.582	4.063	16.880	0.000
Gender(Female)	-5.488	5.454	-1.006	0.316
F508(heterozygos)	3.590	5.503	0.652	0.515
F508(none)	8.971	8.040	1.116	0.266
Age	-1.681	0.379	-4.434	0.000

Pseudoa(Yes)	-2.832	1.024	-2.766	0.006
Gender(Female) * F508(heterozygos)	-2.294	7.751	-0.296	0.768
Gender(Female) * F508(none)	2.877	12.006	0.240	0.811
Gender(Female) * Age	-0.200	0.529	-0.377	0.706
F508(heterozygos) * Age	0.594	0.525	1.131	0.258
F508(none) * Age	1.336	0.793	1.684	0.093
Gender(Female) * F508(heterozygos) * age	-0.608	0.750	-0.810	0.418
Gender(Female) * F508(none) * age	-0.974	1.193	-0.816	0.414

Model 3

$$Y_{ij} \sim N(U_{i1} + V_i(Age) + x_{ij}\beta, \tau^2)$$

$$U_i \sim N(0, \sigma^2)$$

$$Cov(V_i(Age + h), V_i(Age)) = \sigma^2 * exp(-|h|)$$

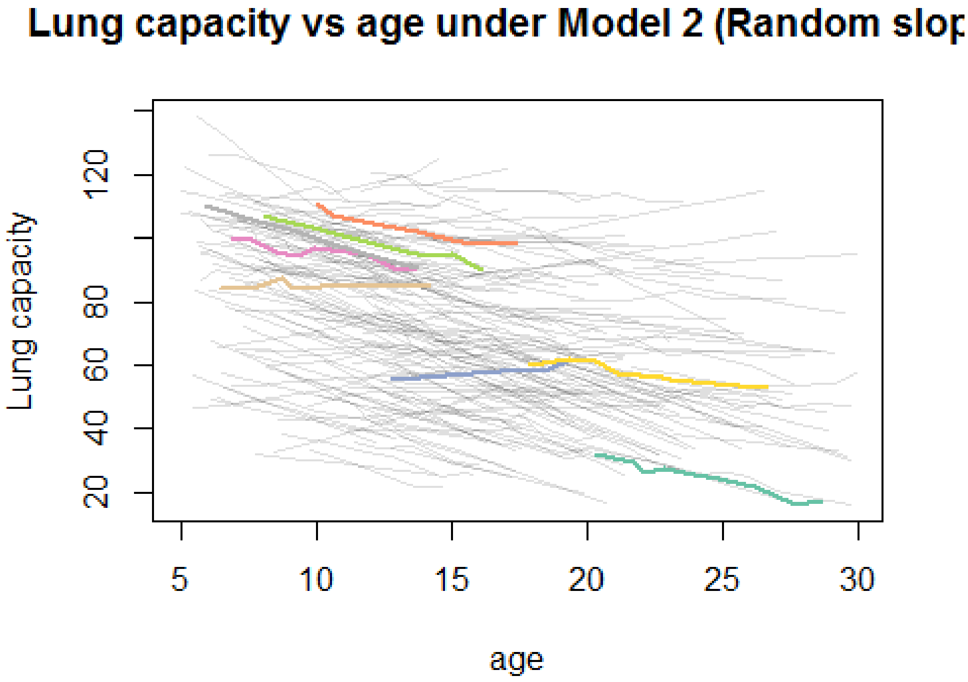
The serial correlation with age represents a contant correlation between the lung function at current age and at later age within each subject. Result of the model 3 is shown below.

Model 3 - Temporal Correlation

	MLE	Std.Error	DF	t-value	p-value
Intercept	66.0733	3.8418	1306	17.1983	0.0000
Gender(Female)	-0.3943	5.0637	194	-0.0779	0.9380
F508(heterozygos)	7.9315	5.0409	194	1.5734	0.1172
F508(none)	8.8664	7.2863	194	1.2169	0.2251
Age	-2.0441	0.3626	1306	-5.6372	0.0000
Pseudoa(Yes)	-2.9885	0.9977	1306	-2.9954	0.0028
Gender(Female) * F508(heterozygos)	-8.3275	7.0197	194	-1.1863	0.2370
Gender(Female) * F508(none)	0.6899	11.0483	194	0.0624	0.9503
Gender(Female) * Age	0.3371	0.5054	1306	0.6670	0.5049
F508(heterozygos) * Age	0.9776	0.4852	1306	2.0148	0.0441
F508(none) * Age	1.4425	0.7503	1306	1.9224	0.0548
Gender(Female) * F508(heterozygos) * age	-1.0388	0.6977	1306	-1.4889	0.1367
Gender(Female) * F508(none) * age	-1.4054	1.1510	1306	-1.2210	0.2223
σ_U	19.0924	NA	NA	NA	NA
τ	9.9379	NA	NA	NA	NA
range	5.9620	NA	NA	NA	NA
σ_V	13.2945	NA	NA	NA	NA

discussion on differences of models

The difference between model 1 and mode 2 is that model 2 accounts for different slope in age for each subject in addtion to model 1. Model 3 describes that within same subject, the influence of age depends on previous ages and different subjects have different correlations. Model 3 has the strongest assumptions.



reasoning for choosing random slope model

Among all 3 models, I believe the model with random slope (model 2) is the best fit model. As per discussion previously, a different slope per subject should be needed. Model 2 is preferred over model 3 due to below reasoning. Even though we can see significant size of the sigma V, which suggests a strong correlation between ages, the model may not be valid. It is due to the fact that, within each subject, there are not enough obervations to support the model and the lack of observation can cause the high variance. Also, if we look at the plot above, having random slope effect may be sufficient to fit the data well and avoid making more complicated assumptions under model 3.

conclusion on research hypotheses

Now let us put our focus back on the two research questions. From result of model 2, by adding the random slope in age, we could see the estimates of gene's interaction terms with age and gender have become weaker and their P values are above 5%. In other words, there is not enough evident to reject the rate at which lung function declines for CF patients does not differ by F508 gene or the effect of the F508 gene on lung function decline differs for females and males over age respectively. Thus I would come to a completely different conclusion than the medical scientist.

Q3

First Task

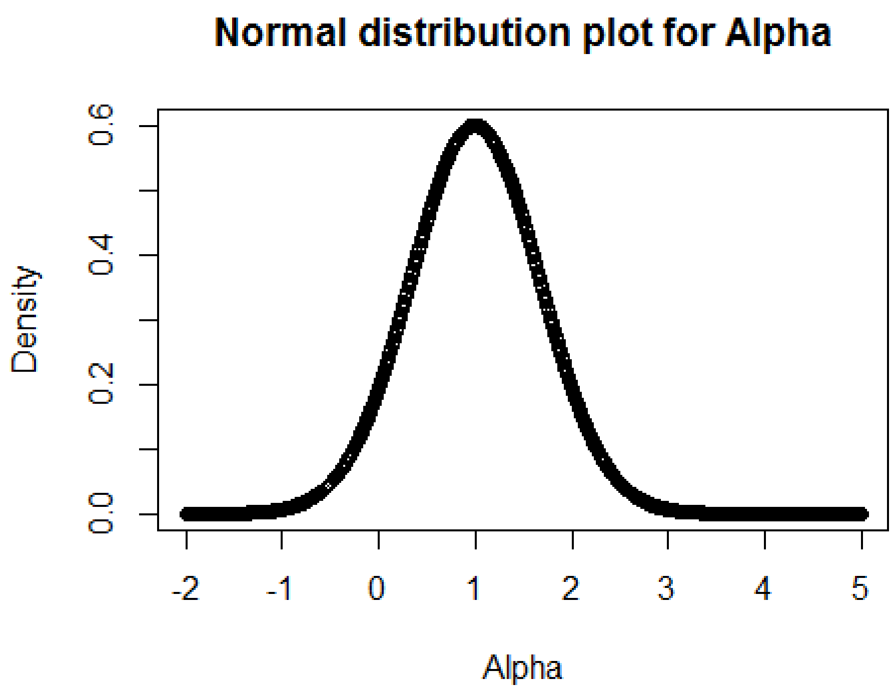
This report is to identify whether the tobacco control programs should target the states or particular school and to conclude if first cigarette smoking has a flat hazard function. First of all, we will discuss how the model is defined basing on the prior infomation given from collaborating scientists.

$$
\begin{aligned}
Y_{ijk} &\sim Weibull(\rho_{ijk}, k) \\
\rho_{ijk} &= exp(-n_{ijk}) \\
n_{ijk} &= x_{ijk}\beta + \mu_i + V_{ij} \\
\mu_i &\sim N(0, \sigma_U^2) \\
V_i &\sim N(0, \sigma_V^2)
\end{aligned}
$$

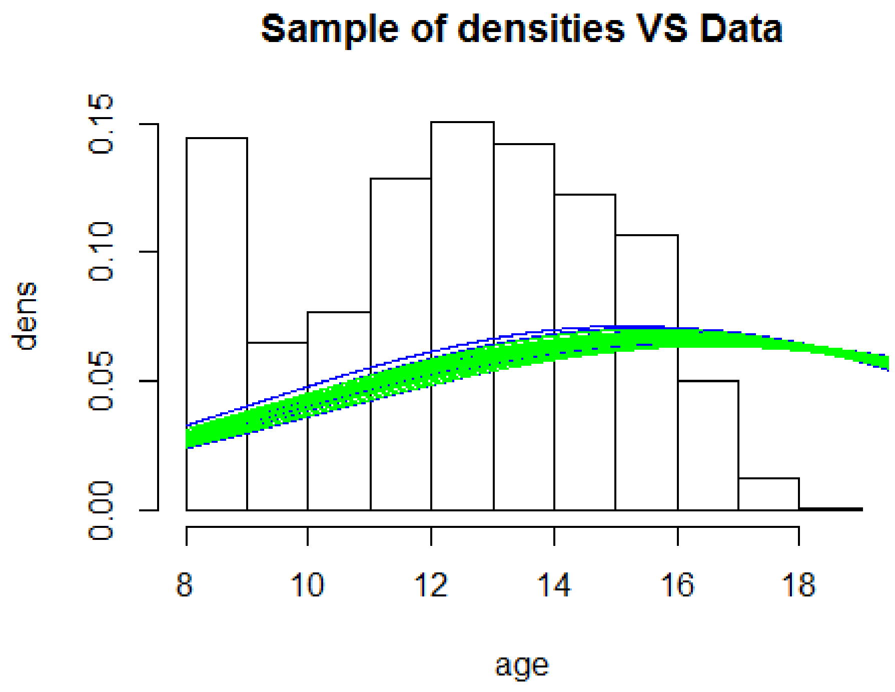
We model the age of first cigarette with Weibull distribution. And we have states and school as random effects. Based on the prior information, since the variation in the rate of smoking initiation between states is not expected to have 5 times. Also the probability that describes all 'unlikely' events is 10%. Therefore penalized complexity prior method is used and I set that there is only 10% of chance to see the the rates lie outside 0.2 to 5. In other words the probability of sigma for state larger than 0.969 is 10%. In a similar way, the probability of sigma for school larger than 0.244 is 10%. In addition, since the flat hazard function is assumed, it is the same way of saying it has an exponential function and alpha equals 1. Thus prior for the alpha has a normal distribution with mean 1 and standard deviation 0.66. Below are the 5% and 95% quantile for school and state standard deviation and a normal distribution plot for alpha.

90% interval for sigma

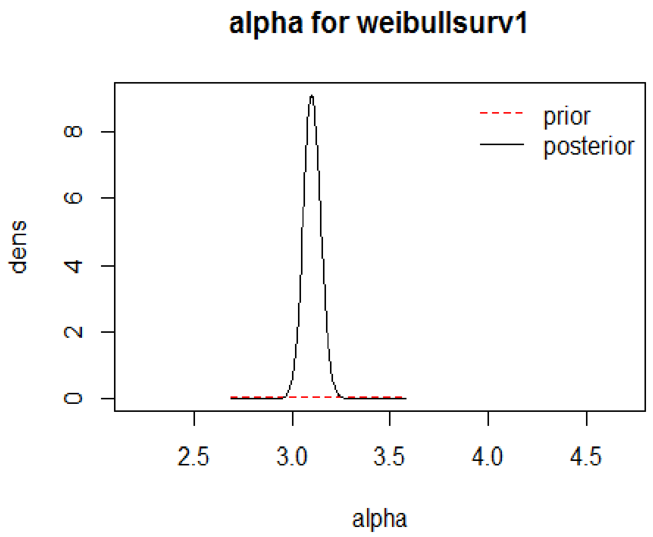
	school	state
0.05	0.6667699	0.1998543
0.95	1.4997677	5.0036458

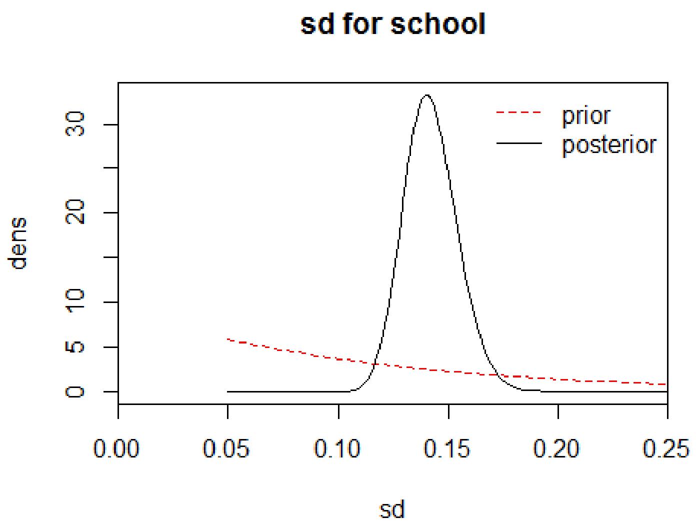


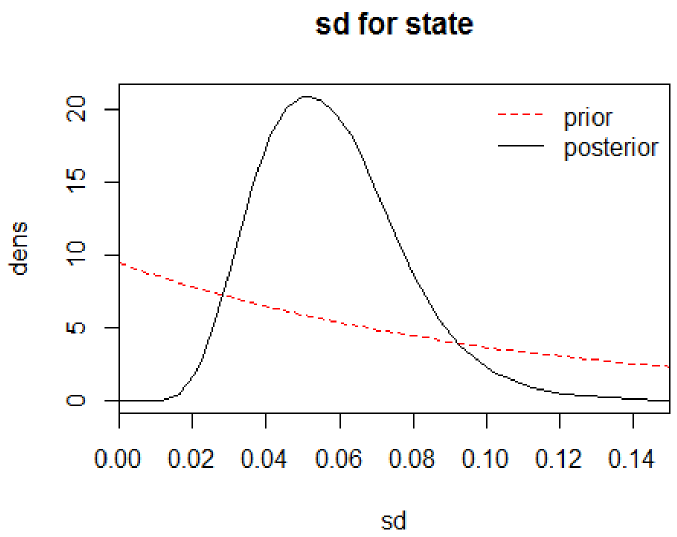
Below is a histogram plot with 2 different INLA models. One is described above (blue lines) and the other is the "default" model (green lines). The default model has penalized complexity on both priors of standard deviation larger 0.5 is around 5%. Also the prior of alpha is a normal with mean 4. By looking at the plot those two models do not differ by much. But since the model above (blue lines) have more proper prior settings so we will continue the analysis with it.



After running the model, we need to investigate the 2 hypotheses. Let us look at the plot below. Firstly, comparing the postieror values of standard deviations between schools and states, we can see that the variantion between school is much larger than between states. Therefore tobacco control programs should target the “worst” school instead. Secondly, the postieror of alpha spikes at around 3.2 and the probability of reaching 1 is extremely low. Therefore we could conclude the hazard function for smoke initiation age is not likely to be flat.







Secondary Task

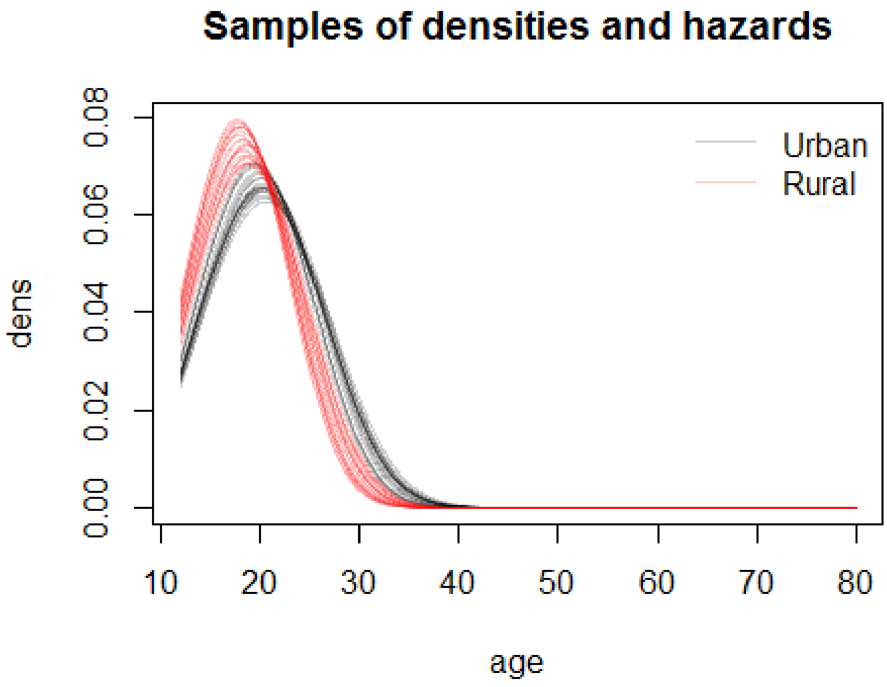
The next task is to discuss what may first smoking ages be expected if a large number of white urban or rural males were obtained. We first look at the estimates below in natural scale.

	mean	0.025quant	0.975quant
Intercept	0.5506925	0.5222060	0.5811746
RuralUrban(Rural)	1.1146057	1.0529975	1.1794411
Sex(Female)	0.9517874	0.9260327	0.9781094
Race(Black)	0.9536436	0.9149863	0.9932951

Race(Hispanic)	1.0270653	0.9933779	1.0616865
Race(Asian)	0.8312481	0.7609764	0.9031797
Race(Native)	1.1137926	1.0061072	1.2242573
Race(Pacific)	1.1854927	1.0089292	1.3684739
Sex(Female) * Race(Black)	0.9857171	0.9328987	1.0413881
Sex(Female) * Race(Hispanic)	1.0197046	0.9755450	1.0658486
Sex(Female) * Race(Asian)	1.0013283	0.8858139	1.1308634
Sex(Female) * Race(Native)	0.9597169	0.8248813	1.1130629
Sex(Female) * Race(Pacific)	0.8495281	0.6172956	1.1272448
SD for school	1.1526924	1.1273298	1.1825963
SD for state	1.0593002	1.0265522	1.1068526

We could see that holding other covariates constant and having white people as the reference group, people who are living in rural, male, hipanic or native or pacific would have earlier age of starting first smoke. It is worhly to note that living in rural area seems to start smoking at earlier age, and this may be helpful obervation when we discuss the expectation of first smking ages between white urban and rural males. It is also worth noticing that different sexes in each different race would have almost similar influnce on expectance of smoke initiation since their magnitudes are all small and not statistically significant.

Now let us discuss the trends of smoking age initiation for urban and rural white males from the sample densities and hazards plot below. We could see rural white males are more likely to start first smoking at earlier age. The highest chance for them to start is around 19 to 20 years old. Meanwhile the highest chance for urban white males to start their first smoke is around 22 to 23 years old.



Appendix

~~~~
---
title: "Sta2201 Assignment 2"
author: "Jiahui Du"
date: "Feb 16, 2019"
output:
  word_document: default
  html_document: default
  pdf_document: default
---
##### Student number: 998268556

#Q1
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE )

library('ggplot2')
library('nlme')
library("lme4")
library("epiDisplay")

data("MathAchieve", package = "MEMSS")
MathAchieve_temp = as.data.frame(MathAchieve)
MathAchieve = subset(MathAchieve_temp,MathAchieve_temp$MathAch >= 0 )




MathAchieve$trans_math = MathAchieve$MathAch
```
~~~~

At first we will look at the shape of histogram of the data below in order to fit a proper model for Math score.

~~~~
```{r}
hist(as.numeric(MathAchieve$trans_math), prob=TRUE, main="Histogram of Math score" , xlab ="Math score")
```
~~~~

<br>

The Math score apears to be left skewed and it is always greater than zero. These evidenets may prevent us from using Normal distribution. And gamma may seem to fit the Math score with above mentioned features of the data.

<br>

~~~~
```{r}
#
# table = rbind(
# mean(MathAchieve$trans_math),
# median(MathAchieve$trans_math),
# sqrt(var(MathAchieve$trans_math)),
# quantile(MathAchieve$trans_math,0.25),
# quantile(MathAchieve$trans_math,0.75))
#
# rownames(table) = c("mean", "median", "SD", "0.25 quantile", "0.75 quantile")
# colnames(table) = c("")
#
# knitr::kable(table,digits=3)
#
# boxplot(MathAchieve$trans_math, main="Boxplot of Math score",
#    xlab="", ylab="Math score")
#
#
# qqnorm(MathAchieve$trans_math)
~~~~

~~~
# qqline(MathAchieve$trans_math)
library(fitdistrplus)


gamma_score = as.vector(MathAchieve$trans_math)
gammafit  <-  fitdistrplus::fitdist(gamma_score, "gamma")
qqcomp(gammafit,main="QQ plot of Math score with Gamma Distribution"  )


###################fit with gamma
# test = glm(trans_math ~ Minority + SES + School, family = Gamma(link = "log"),  data=MathAchieve)
#
#
#  shape = 1/summary(test)$dispersion
#  scale = exp(test$coef["(Intercept)"])/shape



test_re = glmer(trans_math ~ Minority + SES + (1|School), family = Gamma(link = "log"), data=MathAchieve)


 shape = 1/sigma(test_re)^2
 scale = exp(2.59446)/shape


hist(as.numeric(MathAchieve$trans_math), prob=TRUE, main="Histogram of Math score with Gamma Distribution" , xlab ="Math score",ylim=c(0,0.1))
xSeq = seq(0, 120, len = 1000)
lines(xSeq, dgamma(xSeq, shape = shape, scale = scale), col = "red")



```
~~~

<br>

After fitting the Gamma, a QQ-plot of gamma is shown above. We could see the model fit weill until when it reaches 25. It is because there is a maximum math score cut off at 25. From the histogram we could see the gamma model over weights around math score at 10.

<br>

In order to find out whether the differences between schools are greater than within-school variation, a model with school as random effect will be used. The model will give shape and scale values for the gamma thence we get the variance of test score for a typical school test score. Then we simulate the random intercept from a normal distribution with 0 mean and 0.103 standard deviation (from VarCorr(model)) and use those simulated random intercept to simulate the test score. And we could find the sampled variance to estimate the variance of school average test score.

<br>

~~~
```{r}

#var typical school
up = qgamma(0.975, shape=shape, scale=scale)
low = qgamma(0.025, shape=shape, scale=scale)

#var school avg
u = rnorm(1000, mean = 0, sd = 0.103)
sim_y = exp(coef(summary(test_re))[1,1] + u)

low2 = quantile(sim_y , 0.025)
up2 = quantile(sim_y , 0.975)


result = rbind(c(low,up), c(low2,up2))
colnames(result) = c("2.5%", "97.5%" )
rownames(result) = c("Typical School","School Average")

knitr::kable(result, digits=3)
#, escape = FALSE, format = "html", caption = 'Variance of Random Effects')
~~~

~~~
```
~~~

<br>

From the table above, we could see the whole 95% gamma quantile interval for school average is contained within the typical school's. Therefore we could conclude the within school variation is much bigger than the between school variation.

#Q2

~~~
```{r results=FALSE, comment=FALSE, include=FALSE, quietly = TRUE}
#use slide lmmDependent(serial correlation) and lmm(random slope)
dataDir = "C:\\Users\\EDDY\\Documents\\UNIVERSITY\\STA2201_AS\\Hwk\\2\\"
CF = file.path(dataDir, "CF.Rdata")

if (!file.exists(CF)) {
download.file("http://pbrown.ca/teaching/astwo/data/CF.RData", smokeFile)
}
(load(CF))
```
~~~

This is to study the effect of the F508 gene on the decline in lung function in individuals with cystic fibrosis over age. There are mainly two research hypotheses: the rate at which lung function declines for

CF patients depends on the F508 gene and the effect of the F508 gene on lung function decline differs for females and males over age.

<br>

```{r]}
#question: interaction for model2 & 3? yes
#question: how to write in math notation? done
```

A model (Model 1) was suggested bt the medical scientist and it can be represented as:

<br>

###Model 1

$$
\begin{aligned}
Y_{ij} \sim N(U_{i} + x_{ij}\beta , \tau^2)\\
U_{i} \sim N(0, \sigma^2)
\end{aligned}
$$

where Ui is the random intercept for each subject ID. And the random intercept means that each subject may have different lung function. Also, the interaction term between Age, F508 and gender means the effect of one of the terms to lung function differs per different values of the other two terms. Similar idea applies to the two term interactions. And these interactions will help us answer the research hypotheses. Below are the results medical scientist created.

```{r}
library("nlme")
x$ageC = x$AGE - 18
resS = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~1 |ID, data = x)
```

~~~
names(resS$coefficients$fixed) = c('Intercept', 'Gender(Female)', 'F508(heterozygos) ', 'F508(none)', 'Age', 'Pseudoa(Yes)', 'Gender(Female) * F508(heterozygos)', 'Gender(Female) * F508(none)', 'Gender(Female) * Age', 'F508(heterozygos) * Age', 'F508(none) * Age', 'Gender(Female) * F508(heterozygos) * age',  'Gender(Female) * F508(none) * age')


knitr::kable(summary(resS)$tTable, digits = 3)

#escape = FALSE, format = "html", caption  = 'Model 1 - Random intercept',


```
~~~

<br>

~~~
```{r}
#question: important features same as previous question? should i include plots? use default plot and talk about random slope is needed

Scol = rep_len(RColorBrewer::brewer.pal(12, 'Set3'),nlevels(x$F508))

names(Scol) = levels(x$F508)


plot(x$AGE, x$FEV1, type = "n", xlab = "Age", ylab = "Lung capacity", main='Plot of Age VS Lung Capacity per subject')

junk = by(x, x$ID, function(qq) {

lines(qq$AGE, qq$FEV1, col = Scol[as.character(qq$F508)])})

legend("topright", lty = 1, col = Scol, legend = names(Scol), bty = "n")
```
~~~

<br>

However, if we look at the plot above, the slope between age and lung capacity is not necessary to be the same given same gene and age for each subject. This may suggest a random slope in age or a correlation of age within same subject is needed. Therefore, there are two additional models that now we are considering. They are a model with random slope for age (model 2) and a model with serial correlation of lung capacity over age (model 3) and they are represented as below:

<br>

###Model 2

$$
\begin{aligned}
Y_{ij} \sim N(U_{i1} + U_{i2}Age + x_{ij}\beta , \tau^2)\\
\begin{pmatrix} U_{i1}\\U_{i2} \end{pmatrix} \sim MVN(0, \Gamma)
\end{aligned}
$$

The random slop for age from Model 2 indicates that there is a different rate of change in lung capacity as age increases per subject. Below is the result.

~~~
```{r}

#fitting random slope

cf_rs = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~1 + ageC |ID, data = x)


# cf_rs_2int = lme(FEV1 ~ GENDER * F508 + F508 * ageC + GENDER*ageC + PSEUDOA, random = ~1 + ageC |ID, data = x)

# cf_rs_2int2 = lme(FEV1 ~  GENDER * F508 +  GENDER*ageC + PSEUDOA, random = ~1 + ageC |ID, data = x)

# cf_rs_2int3 = lme(FEV1 ~  GENDER * F508 +  ageC + PSEUDOA, random = ~1 + ageC |ID, data = x)

# cf_rs_2int4 = lme(FEV1 ~  GENDER + F508 +  ageC + PSEUDOA, random = ~1 + ageC |ID, data = x)


#summary(cf_rs)




names(cf_rs$coefficients$fixed) = c('Intercept', 'Gender(Female)', 'F508(heterozygos) ', 'F508(none)', 'Age', 'Pseudoa(Yes)', 'Gender(Female) * F508(heterozygos)', 'Gender(Female) * F508(none)', 'Gender(Female) * Age', 'F508(heterozygos) * Age', 'F508(none) * Age', 'Gender(Female) * F508(heterozygos) * age',  'Gender(Female) * F508(none) * age')


theTable = summary(cf_rs)$tTable[, -3]

knitr::kable( theTable[grep("^t[1-5]", rownames(theTable), invert = TRUE), ], digits = 3)

#escape = FALSE, format = "html", caption ='Model 2 - Random Slope',

```
~~~

```{r}
#anova(resS,cf_rs)
# library("lmtest")
# lrtest(cf_rs, cf_rs_2int)
# lrtest(cf_rs_2int, cf_rs_2int2)
# lrtest(cf_rs_2int2, cf_rs_2int3)
# lrtest(cf_rs_2int3, cf_rs_2int4)
```

<br>

###Model 3

$$
\begin{aligned}
Y_{ij} \sim N(U_{i1} + V_{i}(Age) + x_{ij}\beta , \tau^2)\\
U_{i} \sim N(0, \sigma^2)\\
Cov(V_{i}(Age + h), V_{i}(Age)) = \sigma^2*exp(-|h|)
\end{aligned}
$$

The serial correlation with age represents a contant correlation between the lung function at current age and at later age within each subject. Result of the model 3 is shown below.

<br>

```{r}
#fitting serial correlation
cf_sc = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~1|ID, data=x,
correlation=corExp(form=~ageC|ID, nugget=T))
```

~~~
#summary(cf_sc)

names(cf_sc$coefficients$fixed) = c('Intercept', 'Gender(Female)', 'F508(heterozygos) ', 'F508(none)', 'Age', 'Pseudoa(Yes)', 'Gender(Female) * F508(heterozygos)', 'Gender(Female) * F508(none)', 'Gender(Female) * Age', 'F508(heterozygos) * Age', 'F508(none) * Age', 'Gender(Female) * F508(heterozygos) * age',  'Gender(Female) * F508(none) * age')

knitr::kable(Pmisc::lmeTable(cf_sc), digits = 4)
#escape = FALSE, format = "html", caption ='Model 3 - Temporal Correlation',

#question (solved): what is range? Immdependence page 13
```

```{r}
#talk about the difference in fomula, after running, talk about the chg in betas and random effect structure
#Question: do i use log-likehood to compare the 2 only interaction with no interaction to check if significant? try ANOVA
```

<br>

###discussion on differences of models

The difference between model 1 and mode 2 is that model 2 accounts for different slope in age for each subject in addtion to model 1. Model 3 describes that within same subject, the influence of age depends on previous ages and different subjects have different correlations. Model 3 has the strongest assumptions.

<br>

```{r}
CF_plot = data.frame( x=x$AGE,  y=cf_rs$fitted[,'ID'],  id=x$ID)
~~~

```
 S_id = sample(unique(x$ID),8)

 names(S_id) = RColorBrewer::brewer.pal( length(S_id),"Set2")


 plot(CF_plot$x, CF_plot$y, xlab='age', ylab='Lung capacity', type='n', main = 'Lung capacity vs age under Model 2 (Random slope)')

 invisible(by(CF_plot, x$ID, lines, col='#00000020'))


 for(D in 1:length(S_id))

 lines(CF_plot[

 x$ID == S_id[D],c('x','y')],

 col = names(S_id)[D], lwd=2)
```

<br>

###reasoning for choosing random slope model

Among all 3 models, I believe the model with random slope (model 2) is the best fit model. As per discussion previously, a different slope per subject should be needed. Model 2 is preferred over model 3 due to below reasoning. Even though we can see significant size of the sigma V, which suggests a strong correlation between ages, the model may not be valid. It is due to the fact that, within each subject, there are not enough obervations to support the model and the lack of observation can cause the high variance. Also, if we look at the plot above, having random slope effect may be sufficient to fit the data well rather than making more complicated assumpitons under model 3.

<br>

###conclusion on research hypotheses

Now let us put our focus back on the two research questions. From result of model 2, by adding the random slope in age, we could see the estimates of gene's interaction terms with age and gender have become weaker and their P values are above 5%. In other words, there is not enough evident to reject the rate at which lung function declines for CF patients deos not differ by F508 gene or the effect of the F508 gene on lung function decline differs for females and males over age respectively. Thus I would come to a completely different conclusion than the medical scientist.

~~~
<br>

<br>



#Q3

```{r results=FALSE, comment=FALSE, include=FALSE, quietly = TRUE}

dataDir = "C:\\Users\\EDDY\\Documents\\UNIVERSITY\\STA2201_AS\\Hwk\\1\\"

smokeFile = file.path(dataDir, "smokeDownload.RData")

if (!file.exists(smokeFile)) {

download.file("http://pbrown.ca/teaching/astwo/data/smoke.RData", smokeFile)

}

(load(smokeFile))


forInla = smoke[, c("Age", "Age_first_tried_cigt_smkg", "Sex", "Race", "state", "school", "RuralUrban")]

forInla = na.omit(forInla)

forInla = as.list(forInla)

library("INLA")

```



```{r}

#Survive page 40-41,64 plot posterior&prior for SD

#Survive page 11 plot hazard

#Survive page 15 plot model fit
~~~

#Survive page 67 for similar study

#question: compare SD for school and state? yes

#question: plot hazard function and check if its flat? yes and fit with exponential instead of weibull (check scale = 1 (exp) or not)

#question: the given code has tranformed the data already? no, do exp on Betas

#question: uptake = prediction? morebayes page 50

#question: how to state prior?

~~~

###First Task

This report is to identify whether the tobacco control programs should target the states or particular school and to conclude if first cigarette smoking has a flat hazard function. First of all, we will discuss how the model is defined basing on the prior infomation given from collaborating scientists.

$$

\begin{aligned}

Y_{ijk} \sim Weibull(\rho_{ijk}, k)\\

\rho_{ijk} = exp(-n_{ijk})\\

n_{ijk} = x_{ijk}\beta + \mu_{i} + V_{ij}\\

\mu_{i} \sim N(0, \sigma^2_{U})\\

V_{i} \sim N(0, \sigma^2_{V})

\end{aligned}

$$

```{r}

#to run and see what is the best sigma value for prior

# xSeq = seq(0, 2, len = 1500)

# error_min = 1

#
~~~

```
for (i in 1:length(xSeq)) {
result = exp(c(-1.66,1.66)*xSeq[i])
value1 = result[1]
value2 = result[2]
error_now = abs(value1-0.2)+abs(value2-5)
if (error_now<error_min) {error_min = error_now
sigma = xSeq[i]
}
#
}
###############

forSurv = data.frame(time = (pmin(forInla$Age_first_tried_cigt_smkg, forInla$Age) - 4)/10,
event=forInla$Age_first_tried_cigt_smkg <=forInla$Age)

left censoring
forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2
forInla$y = inla.surv(forSurv$time, forSurv$event)

fitS2 = inla(y ~ RuralUrban + Sex * Race
 + f(school,model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(0.2441628,0.1))))
 + f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(0.96998,0.1)))),
control.family = list(variant = 1, hyper = list(alpha = list(prior = "normal", param = c(log(1),(2/3)^(-
2))))),
data = forInla, family = "weibullsurv", control.compute=list(config = TRUE))

fit_d = inla(y ~ RuralUrban + Sex * Race
 + f(school,model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(0.5,0.05))))
 + f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(0.5,0.05)))),
```

control.family = list(variant = 1, hyper = list(alpha = list(prior = "normal", param = c(log(4),(2/3)^(-2))))),

data = forInla, family = "weibullsurv", control.compute=list(config = TRUE))

<br>

We model the age of first cigarette with Weibull distribution. And we have states and school as random effects. Based on the prior information, since the variation in the rate of smoking initiation between states is not expected to have 5 times. Also the probability that describes all 'unlikely' events is 10%. Therefore penalized complexity prior method is used and I set that there is only 10% of chance to see the the rates lie outside 0.2 to 5. In other words the probability of sigma for state larger than 0.969 is 10%. In a similar way, the probability of sigma for school larger than 0.244 is 10%. In addition, since the flat hazard function is assumed, the prior for the alpha has a normal distribution with mean 1 and standard deviation 0.66. Below are the 5% and 95% quantile for school and state standard deviation and a normal distribution plot for alpha.

```{r}

#plot/show prior

#school + state

result = cbind.data.frame(exp(c(-1.66,1.66)*0.2441628), exp(c(-1.66,1.66)*0.96998))

colnames(result)  = c("school","state")

rownames(result) = c("0.05", "0.95")

knitr::kable(result)

#, digits=4, escape = FALSE, format = "html", caption ='90% interval for sigma'

```

```{r}

#plot/show prior

#alpha

xSeq_a = seq(-2, 5, len = 1000)

plot(xSeq_a, dnorm(xSeq_a, mean =  1, sd = 0.666), col = "black", ylab = "Density", xlab = 'Alpha', main = 'Normal distribution plot for Alpha')
```

<br>

Below is a histogram plot with 2 different INLA models. One is described above (blue lines) and the other is the "default" model (green lines). The default model has penalized complexity on both priors of standard deviation larger 0.5 is around 5%. Also the prior of alpha is a normal with mean 4. By looking at the plot those two models do not differ by much. But since the one above (blue lines) have more proper prior settings so we will continue the analysis with it.

<br>

```{r}
#assess model fit


 hist(forInla$Age_first_tried_cigt_smkg, main='Sample of densities VS Data', xlab='age', ylab='dens', prob=TRUE)



 kappa = fitS2$summary.hyperpar['alpha','mode']

 lambda = exp(fitS2$summary.fixed['(Intercept)' , "mode"])

 xSeq = seq(8,20,len=1000)


 densHaz = Pmisc::sampleDensHaz(fit = fitS2, x = xSeq, n = 20, scale = 10)

 matlines(xSeq, densHaz[, "dens", ], type = "l", lty = 1, col = "blue")


  densHaz = Pmisc::sampleDensHaz(fit = fit_d, x = xSeq, n = 20, scale = 10)

  matlines(xSeq, densHaz[, "dens", ], type = "l", lty = 1, col = "green")



 # hazEst = survfit(Surv(forSurv$time, forSurv$event) ~ 1, data=forSurv)
```

~~~
 # plot(hazEst, fun='cumhaz', log='y', xlab='age', ylab = 'cum haz' )

 #

 # matlines(xSeq, densHaz[, "cumhaz", ], type = "l", lty = 1, col = "#FF000020")


```
~~~

<br>

After running the model, we need to investigate the 2 hypotheses. Let us look at the plot below. Firstly, comparing the postieror values of standard deviations between schools and states, we can see that the variantion between school is much larger than between states. Therefore tobacco control programs should target the "worst" school instead. Secondly, the postieror of alpha spikes at around 3.2 and the probability of reaching 1 is extremely low. Therefore we could conclude the hazard function for smoke initiation age is not likely to be flat.

<br>

~~~
```{r}


fitS2$priorPost = Pmisc::priorPost(fitS2)



for (Dparam in fitS2$priorPost$parameters) {

do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)

do.call(legend, fitS2$priorPost$legend)

title(main = Dparam)}
```
~~~

###Secondary Task

<br>

The next task is to discuss what may first smoking ages be expected if a large number of white urban or rural males were obtained. We first look at the estimates below in natural scale.

<br>

```{r}

rownames(fitS2$summary.fixed) = c('Intercept', 'RuralUrban(Rural)', 'Sex(Female)', 'Race(Black)', 'Race(Hispanic)', 'Race(Asian)','Race(Native)',             'Race(Pacific)', 'Sex(Female) * Race(Black)','Sex(Female) * Race(Hispanic)','Sex(Female) * Race(Asian)','Sex(Female) * Race(Native)','Sex(Female) * Race(Pacific)')

est = rbind(exp(fitS2$summary.fixed[, c("mean", "0.025quant","0.975quant")]),

exp(Pmisc::priorPostSd(fitS2)$summary[,c("mean", "0.025quant", "0.975quant")]))

knitr::kable(est)

#, digits=4, escape = FALSE, format = "html", caption ='Results in Natural Scale'

```

<br>

We could see that holding other covariates constant, people who are living in rural, male, hipanic or native or pacific would have earlier age of starting first smoke comparing to white people. It is worhly to note that living in rural area seems to start smoking at earlier age, and this may be helpful obervation when we discuss the expectation of first smking ages between white urban and rural males. It is also worth noticing that different sexes in each different race would have almost similar influnce on expectance of smoke initiation since their magnitudes are all small and not statistically significant.

<br>

Now let us discuss the trends of smoking age initiation for urban and rural white males from the sample densities and hazards plot below. We could see rural white males are more likely to start first smoking at earlier age. The highest chance for them to start is around 19 to 20 years old. Meanwhile the highest chance for urban white males to start their first smoke is around 22 to 23 years old.

<br>

```{r}
```

```
#survival p54
#check
library(mapmisc)
test = as.data.frame(cbind(forInla$RuralUrban,as.factor(forInla$RuralUrban)))

xSeqNatural = seq(12, 80, len=1000)

xSeqTrans = (xSeqNatural - 4)

 newCov = model.matrix(~RuralUrban, data.frame(RuralUrban=unique(forInla$RuralUrban)))

 rownames(newCov) = as.character(unique(forInla$RuralUrban))

 densHaz = Pmisc::sampleDensHaz( fit=fitS2, x=xSeqTrans, n=20, covariates = newCov, scale=10)
 Scol = mapmisc::col2html(1:2, 0.2)
 names(Scol) = c('Urban','Rural')

 plot(NA, xlim = c(12, 80), ylim = range(densHaz[, names(Scol), "dens", ]), xlab = "age", ylab = "dens",
main = 'Samples of densities and hazards')
 legend("topright", lty = 1, col = Scol, legend = names(Scol), bty = "n")

 for (D in names(Scol)) { matlines(xSeqNatural, densHaz[, D, "dens", ], type = "l", lty = 1, col = Scol[D])}
```