Homework 2, Mixed Effects Models

Jessica Chau

2019-03-09

Math Achievement Scores

The histogram in Figure 1 shows that the data does not appear normal. The red curve is the normal distribution plotted with mean and standard deviation of the dataset. The QQ plot is also plotted and shows the same conclusion as the histogram. As a result, the data is not normal and we cannot assume the data is normal.

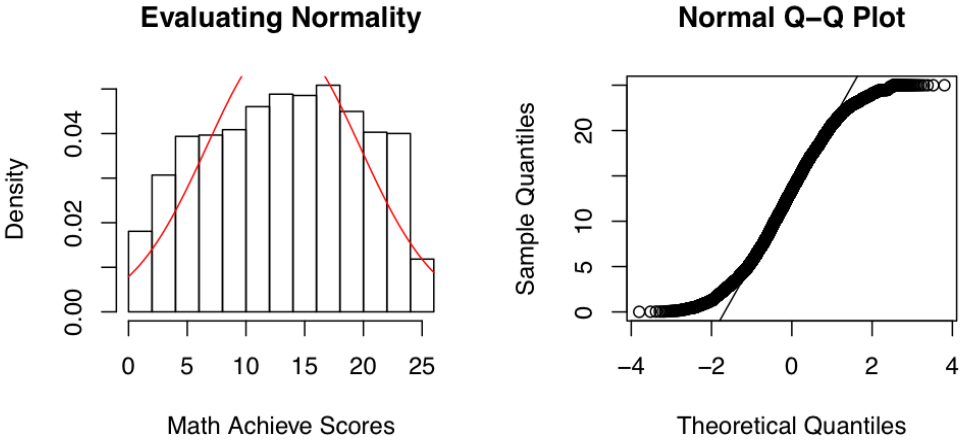

Figure 1: Evaluating Normality of the Math Achieve scores. The Histogram and QQ Plot both show that the data does not appear to be normal

A linear mixed model is selected for this dataset. A linear mixed model is appropriate for datasets with clusteres/nested data such as students in classrooms and repeated-measures experiments, where subjects are repeatedly measured over time or in different conditions. In the Math Achieve dataset, several students' math achievement scores are collected from various schools which satisfies the linear mixed model conditions.

In addtion, linear mixed models are appropriate when the outcome variables' residuals are normally distributed but may not be independent or have constant variance. In Figure 1, the residuals (after fitting the linear mixed models) appear to be normally distributed.

From Table 2, it is observed that the difference within schools is $\tau = 0.103$ whereas the difference between school is $\sigma = 0.479$. However, since the model is not normal, it is difficult to compare τ and σ, thus the 95% interval will be calculated. The 95% interval for test scores of students in the typical school (with random effect of zero) is (3.92, 28.56). With no random effect, the 95% interval is calculated from the gamma distribution after finding the scale and shape using the school with

Table 1: Coefficients of Fixed Effects on Math Scores

	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	2.594	0.017	153.13	0
MinorityYes	-0.244	0.020	-12.11	0
SES	0.166	0.010	15.88	0

Table 2: Variance of School and Residuals

Random effect	Standard Deviation
School	0.103
Residual	0.479

zero random effect. The 95% interval for school average test scores is (10.93, 16.24). It was found that the 95% interval for variation between schools is smaller than variation within schools.

Cystic Fybrosis

In the following analyses, three linear mixed effect models will be scrutinized in order to evaluate the most appropriate model for the F508 dataset. The model terms will be examined and the assumptions will be evaluated for their suitability to the dataset.

Understanding the Models and its Assumptions

The following contains the mathematical notation of the model. Note that there are some repeated notation that will be defined here:

$$\begin{aligned} X_{ij}\beta = {} & (Female)x_1 + (F508heterozygous)x_2 \\ & + (F508none)x_3 + (ageC)x_4 \\ & + (PSEUDOAyes)x_5 + (female : F508heterozygous)x_6 \\ & + (female : F508none)x_7 + (female : ageC)x_8 \\ & + (F508heterozygous : ageC)x_9 + (F508none : ageC)x_{10} \\ & + (female : F508heterozygous : ageC)x_{11} + (female : F508none : ageC)x_{12} \end{aligned} \tag{1}$$

U_i is the ID of the participant in the study.

Model 1: Linear Mixed Effect with Random Intercept Model

$$\begin{aligned} Y_{ij}|U_i &\overset{ind}{\sim} N(\mu_{ij}, \tau^2) \\ \mu_{ij} &= X_{ij}\beta + U_i \\ U_i &\overset{ind}{\sim} N(0, \sigma^2) \end{aligned}$$

Under model 1, the random intercept model, U_i, is the ID of the subject. Each individual's health could be impacted by the effect of the F508 gene differently, a result, ID is treated as a random intercept effect. The linear coefficients in the model, $X_{ij}\beta$, is Pseudoa and the interaction terms: Gender, F508, and age. In this model, it is assumed that for each individual, the outcomes are independent and follows a normal distribution.

The assumptions of Model 1:

- Age has the same effect on the response for each individual
- The errors have constant variance (homogenity of variance)
- The residuals of the model are independent
- The residuals of the model are Normally distributed
- The explanatory variables are related linearly to the response

Model 2: Linear Mixed Effect with Random Slope

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

There are two random effects in this model: intercept and slope. The intercept random effect, U_{i1} is the ID of the individual as explained previously. The slope random effect, U_{i2} is time and varies for each individual. Notice that there are two random effects and together, U_{i1} and U_{i2} follows multivariate normal with mean 0 and variance Γ.

The assumptions of Model 2:

- Observations from the same individual are correlated and requires adjusting for time
- Time covariate impacts each individual differently
- The errors have constant variance (homogenity of variance)
- The residuals of the model are independent
- The residuals of the model are Normally distributed
- The explanatory variables are related linearly to the response

Model 3: Linear Mixed Effect with Serial Correlation

$$Y_{ij}|U, V \overset{ind}{\sim} N(X_{ij}\beta + U_i + V_i(t_{ij}), \tau^2)$$
$$U_i \overset{ind}{\sim} N(0, \sigma^2)$$
$$V \sim MVN(0, \Sigma_V)$$

$$cov[V_i(t+h), V_i(t)] = \sigma^2 exp(\frac{-|h|}{\phi})$$

Note: $cov(V_i(t+h), V_j(t)) = 0$ if $i \neq j$

Similar to model 1 and 2, U_i is the intercept random effect in the model (the ID). $V_i(t)$ is the effect of time for each individual in the study. In model 3, the serial correlation is measured by the $V_i(t)$ random effect and is dependent on time.

The assumptions of Model 3:

- Age has different size effects on the response for each individual
- The errors have constant variance (homogenity of variance)
- The residuals of the model are independent
- The residuals of the model are Normally distributed
- The explanatory variables are related linearly to the response

Table 3: Random Intercept Model Summary of Covariates on Lung Function

	Value	Std.Error	DF	t-value	p-value
(Intercept)	66.877	3.749	1306	17.840	0.000
GENDERfemale	-2.203	4.996	194	-0.441	0.660
F508heterozygous	6.269	5.001	194	1.254	0.212
F508none	7.157	7.348	194	0.974	0.331
ageC	-1.754	0.251	1306	-6.989	0.000
PSEUDOAyes	-2.159	1.059	1306	-2.038	0.042
GENDERfemale:F508heterozygous	-6.275	7.031	194	-0.892	0.373
GENDERfemale:F508none	2.014	11.032	194	0.183	0.855
GENDERfemale:ageC	0.071	0.352	1306	0.200	0.841
F508heterozygous:ageC	0.744	0.344	1306	2.165	0.031
F508none:ageC	1.610	0.534	1306	3.014	0.003
GENDERfemale:F508heterozygous:ageC	-0.998	0.495	1306	-2.018	0.044
GENDERfemale:F508none:ageC	-1.889	0.810	1306	-2.333	0.020

Differences between the three Models

In model 1, there is a random intercept whereas in model 2, there is a random intercept and random slope. In model 1, the random intercept takes into account the baseline differences of the individual but assume that the other effects (gender, F508 and age) will have the same impact on each individual. As shown in Table 3, model 1 does support both research hypotheses.

In model 2, it contains the random slope, time. The random slope assumes that time impacts individuals differently - some individuals are expected to have bigger or smaller effect as time increases. A random slope model allows for differing slope for the effects of time. As shown in Table 4, model 2 does not support either research hypotheses.

In model 3, it assumes that the random effect depends on the age variable. This assumes that there is no correlation between different individuals but there is correlation for the individual over time. This model adjusts for the correlation exhibited on an individual over the time of the study. As shown in Table 5, model 3 does not support the research hypotheses.

Table 4: Random Slope Model Summary of Covariates on Lung Function

	Value	Std.Error	DF	t-value	p-value
(Intercept)	68.582	4.063	1306	16.880	0.000
GENDERfemale	-5.488	5.454	194	-1.006	0.316
F508heterozygous	3.590	5.503	194	0.652	0.515
F508none	8.971	8.040	194	1.116	0.266
ageC	-1.681	0.379	1306	-4.434	0.000
PSEUDOAyes	-2.832	1.024	1306	-2.766	0.006
GENDERfemale:F508heterozygous	-2.294	7.751	194	-0.296	0.768
GENDERfemale:F508none	2.877	12.006	194	0.240	0.811
GENDERfemale:ageC	-0.200	0.529	1306	-0.377	0.706
F508heterozygous:ageC	0.594	0.525	1306	1.131	0.258
F508none:ageC	1.336	0.793	1306	1.684	0.093
GENDERfemale:F508heterozygous:ageC	-0.608	0.750	1306	-0.810	0.418
GENDERfemale:F508none:ageC	-0.974	1.193	1306	-0.816	0.414

Table 5: Serial Correlation Model Summary of Covariates on Lung Function

	Value	Std.Error	DF	t-value	p-value
(Intercept)	66.073	3.842	1306	17.198	0.000
GENDERfemale	-0.394	5.064	194	-0.078	0.938
F508heterozygous	7.931	5.041	194	1.573	0.117
F508none	8.866	7.286	194	1.217	0.225
ageC	-2.044	0.363	1306	-5.637	0.000
PSEUDOAyes	-2.988	0.998	1306	-2.995	0.003
GENDERfemale:F508heterozygous	-8.328	7.020	194	-1.186	0.237
GENDERfemale:F508none	0.690	11.048	194	0.062	0.950
GENDERfemale:ageC	0.337	0.505	1306	0.667	0.505
F508heterozygous:ageC	0.978	0.485	1306	2.015	0.044
F508none:ageC	1.442	0.750	1306	1.922	0.055
GENDERfemale:F508heterozygous:ageC	-1.039	0.698	1306	-1.489	0.137
GENDERfemale:F508none:ageC	-1.405	1.151	1306	-1.221	0.222

Model Conclusion

An anova test is performed to compare the performance of the three models. The random slope is the most suitable model. The anova test showed that model 2 and model 3 were more suitable than model 1. Although $V_i(t)$ is fairly large, it may not suggest a correlation because there are very few observations per participant. The maximum number of observations per participant was 9 and the minimum number of observations was 6. In addition, the residuals were plotted and it did not appear to have a trend as time increases. From the summary table in model 2, it can be concluded that both research hypothesis are not supported. The rate at which lung function declines for CF patients does not depend on the F508 gene at the 5% significance level. The effect of the F508 gene on lung function decline does not differ for females and males. The conclusions are different from the medical scientists' because in model 2 takes into consideration the random effect time has on the data. In the medical scientist's model, it assumes time only impact the baseline of the individual.

Table 6: Random Slope Model is statistically significant (pvalue <0.05), which indicates it fits the data better than Random Intercept Model

	df	AIC	p-value
model_random_intercept	15	12504	NA
model_random_slope	17	12412	0

Table 7: Serial Correlation Model is statistically significant (pvalue <0.05), which indicates it fits the data better than Random Intercept Model

	df	AIC	p-value
model_random_intercept	15	12504	NA
CF_time_model	17	12351	0

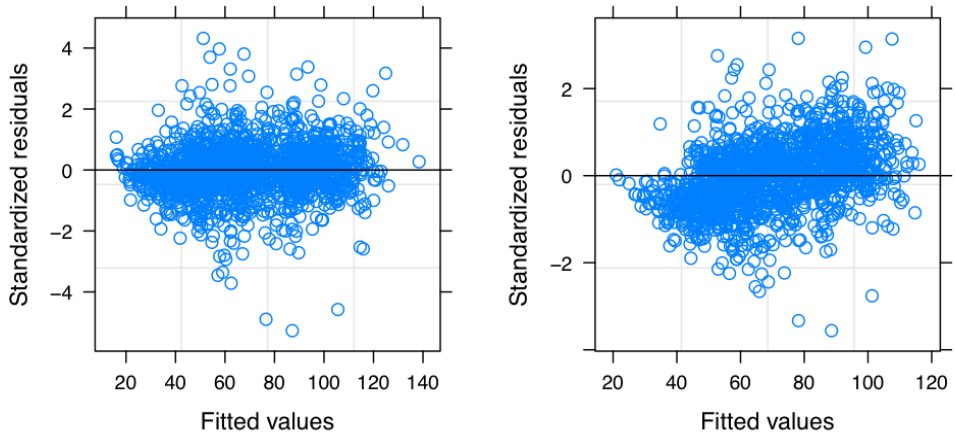

Figure 2: These figures plots the residuals of the random slope and serial correlation models. Residuals from the random slope model does not appear to have serial correlation

Smoking

Introduction

According to the National Health Service, smoking can cause several health issues such as lung cancer, increase risk of heart attacks, stomach cancer and more. As a result, it is important to understand the trends of youths who are starting to smoke in order to prevent it from happening. In addition, by finding demographic trends in smoking, we can also better allocate resources (such a smoking prevention education) to areas that require them.

The first hypothesis determines if geographic variation is larger than school variation on children smoking earlier while controlling for confounding factors. The second hypothesis evaluates if age of smoking has the same probability of trying cigarettes in the next month. In addition to the two hypotheses, this report also explores if there is a difference in smoking uptakes between white males

living in urban areas compared to rural ones.

The data used in this study is from the 2014 American National Youth Tobacco Survey. The survey is given to a sample of public and private school students enrolled in regular middle schools and high schools in the 50 US states.

Methods

To address the hypothesis and secondary problem, it is important to build a model that incorporates the collaboring scientists' prior information. Both models use Bayesian Inference, however the second model restricts the priors.

In both models, there are three parameters with priors to assess in the model: $\alpha, \sigma_1, \sigma_2$

$$Y_i \sim Weibull(\lambda_i, \alpha)$$
$$\lambda_i = e^{-(X\beta + U_{school} + V_{state})}$$
$$U_{school} \sim N(0, \sigma_1^2)$$
$$V_{state} \sim N(0, \sigma_2^2)$$

σ_1 and σ_2 both are parameterized using PC prior for precision. Given the prior information from the scientists, the model is modified. Within a given state, the worst states have 5 or 10 times the rate of the healthiest states. This suggests that λ_1 would be equal to 5 and it would be possible to calculate σ_1 with this. As a result, σ_1 is modelled with penalized complexity precision of (1.794, 0.1). Given that the worst schools have at most 50% greater rate than the 'healthiest' schools, $\lambda_2 = 1.5$ and σ_2 also follows penalized complexity precision with (0.901, 0.1) The flat hazard rate indicates that the mean of the normal distribution is 1. As a result, the α of the weibull survival distribution follows Normal(1,(2/3)ˆ(-2))). The most appropriate model should contain the prior information from the collaboring scientists. The prior vs. posterior graphs shows that the modified model performs well, as indicated by the narrow shape of the posterior graphs. The modified prior model will be used to assess the research hypothesis.

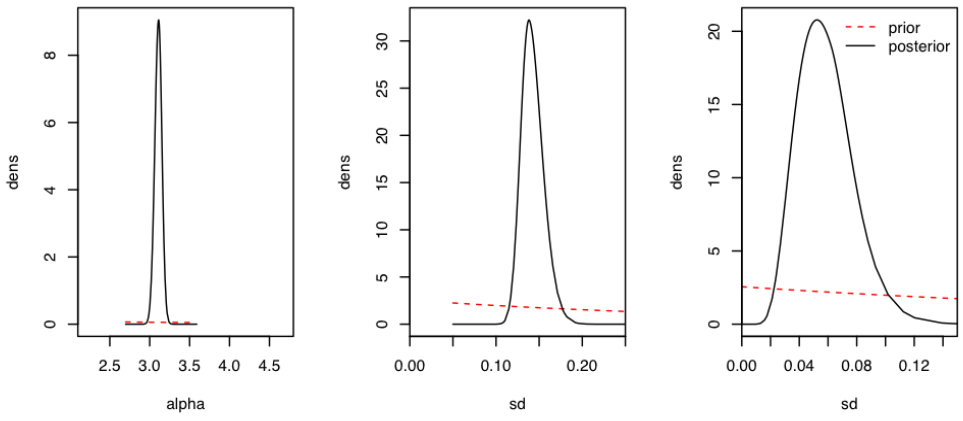

Figure 3: Posterior vs. Prior of Modified Prior Model

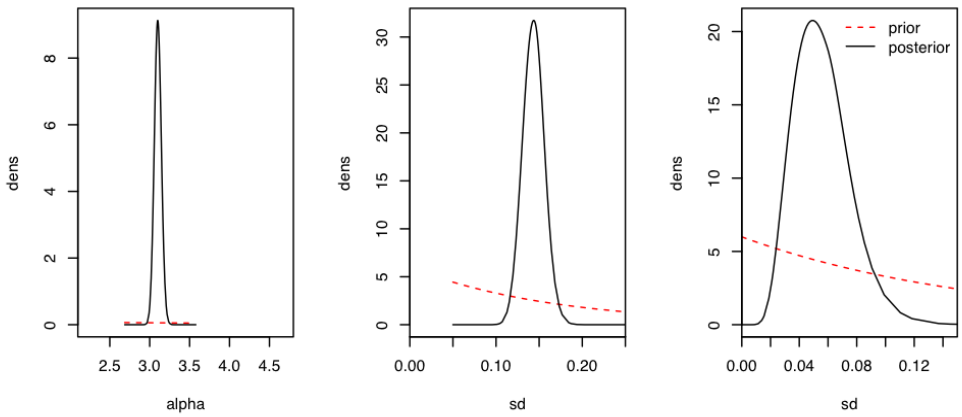

Figure 4: Posterior vs. Prior of Original Model

Results - Research Hypotheses

In the first research question, it tries to understand whether geographic variation is larger than school variation. Table 8 shows that the variation between school (0.143) is larger than the variation between states (0.058). It is suffice to say that it is difficult to target specific states with so much variability between the first smoking ages within the schools. In the second research question, it tries to understand if the hazard function is flat. In figure 3, the posterior of alpha is centered at approximately 3.1, as a result, the likelihood of the hazard being 1 (which indicates a flat hazard) is low. The hazard function for the age that a student starts smoking is unlikely to be flat.

In table 9, it shows the summary of model parameters and its impact on the age that youths first begin smoking. Youths who have the following demographics tends to be more likely to smoke earlier:

- Rural (smoke earlier by 11.4
- Native (smoke earlier by 11.45)
- Hispanic (smoke earlier by 2.7
- Pacific (smoke earlier by 18.9
- Hispanic Females (smoke earlier by 2.0
- Asian Females (smoke earlier by 0.01

Generally, these are the demographics that tends to smoke later:

- Females (smoke later by 4.8
- Black (smoke later by 4.6
- Asian (smoke later by 16.8
- Black Females (smoke later by 1.4

- Native Females (smoke later by 4.0
- Pacific Females (smoke later by 14.5

Table 8: Evaluation of Variation Between State and School

	mean	sd	0.025quant	0.5quant	0.975quant	mode
SD for school	0.143	0.013	0.120	0.141	0.171	0.138
SD for state	0.058	0.019	0.027	0.057	0.102	0.053

Table 9: Summary of Model Parameters

	Estimate	Lower 2.5%	Upper 97.5%
(Intercept)	0.551	0.523	0.582
RuralUrbanRural	1.114	1.053	1.179
SexF	0.952	0.926	0.978
Raceblack	0.954	0.915	0.993
Racehispanic	1.027	0.993	1.062
Raceasian	0.832	0.761	0.903
Racenative	1.115	1.006	1.224
Racepacific	1.189	1.009	1.368
SexF:Raceblack	0.986	0.933	1.041
SexF:Racehispanic	1.020	0.976	1.066
SexF:Raceasian	1.001	0.886	1.131
SexF:Racenative	0.960	0.825	1.113
SexF:Racepacific	0.855	0.618	1.127
alpha parameter for weibullsurv	3.108	3.019	3.192
Precision for school	50.011	34.103	68.983
Precision for state	313.067	96.628	1389.613

Results - Secondary question

To convey the differences between white urban males and white rural males in their smoking uptake habits, a graph that shows the probabilities changing by age is plotted in Figure 5. It appears that the likelihood of rural smoking uptake tends to be higher from age 0 to the early twenties and urban smoking uptake is higher in the mid-twenties to early forties before the probabilities taper off.

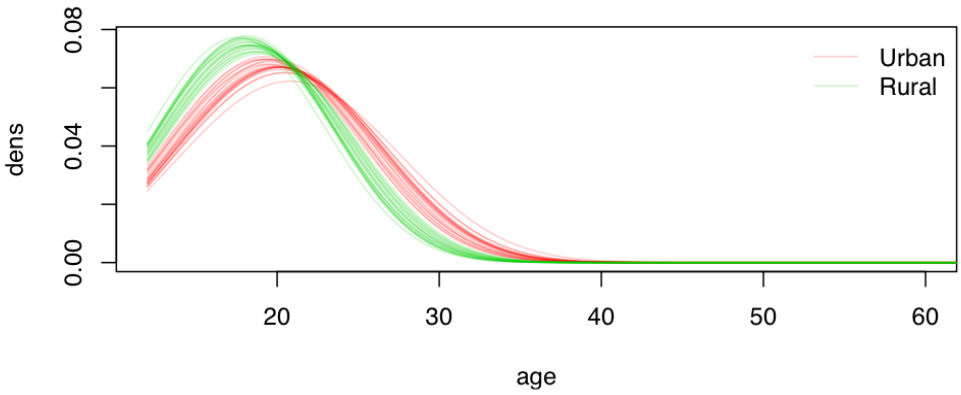

Figure 5: Smoking Uptake Probabilities of Urban vs. Rural White Males

Rcode References

```r
set.seed(0)
data("MathAchieve", package="MEMSS")
# head(MathAchieve)
# help(MathAchieve)
# nrow(MathAchieve) #7185 rows

# Removing negative MathScores
math_dat <- MathAchieve[MathAchieve$MathAch >= 0, ]
#nrow(math_dat) #6973
# summary(math_dat)

# response is left-skewed
#hist(math_dat$MathAch, breaks=50)

# par(mfrow=c(1,2))
# hist(math_dat$MathAch,
#      main="Evaluating Normality", freq=FALSE,
#      xlab="Math Achieve Scores", ylab="Density")
# curve(dnorm(x, mean=mean(math_dat$MathAch), sd=sd(math_dat$MathAch)),  col="red", add=TRUE)
# qqnorm(math_dat$MathAch)
# qqline(math_dat$MathAch)

# model glmr
model_1 = glmer(MathAch~Minority+SES+(1|School), family=Gamma(link="log"), data = math_dat)
# summary(model_1)
# knitr::kable(coef(summary(model_1)), digits = 3, escape = FALSE, format = "latex", caption =

vc <- VarCorr(model_1)
vc_tbl = as.data.frame(vc)
# knitr::kable(vc_tbl[, c(1,5)], digits = 3, escape = FALSE, format = "latex",
#              col.names = c("Random effect", "Standard Deviation"),
#              caption = 'Variance of School and Residuals')

# summary(model_1)
# within school
# Yij ~ GAMMA(XB, b)
intercept = coef(summary(model_1))[ , "Estimate"][1]
shape_est = 1/sigma(model_1)^2 #https://stat.ethz.ch/pipermail/r-sig-mixed-models/2015q1/02337
scale_est = exp(intercept)/shape_est
upper = qgamma(0.975, shape_est, scale = scale_est, lower.tail = TRUE, log.p = FALSE)
lower = qgamma(0.025, shape_est, scale = scale_est, lower.tail = TRUE, log.p = FALSE)
# 3.923 & 28.56


# between schools
```

```r
# summary(model_1)
# simulate the Ui
sim_u = rnorm(1000, mean = 0, sd = 0.103)
expected_sim_y = exp(intercept+sim_u)
lower = quantile(expected_sim_y, 0.025) #10.93
upper = quantile(expected_sim_y, 0.975) #16.24


cUrl = paste("https://faculty.washington.edu/heagerty/Courses/VA-longitudinal/private/",
c("NewCFkids.data", "NewCFkids.txt"), sep = "")
cFile = file.path("..", "data", basename(cUrl))
for (D in 1:length(cFile)) {
if (!file.exists(cFile[D])) {
download.file(cUrl[D], cFile[D])
}
}
x = read.table(cFile[1], header = FALSE)
header = scan(cFile[2], skip = 11, n = ncol(x), what = "a",
sep = "\n")
header = matrix(trimws(unlist(strsplit(header, " = "))),
ncol = 2, byrow = TRUE)
colnames(x) = header[, 1]
factors = header[grep("=", header[, 2]), ]

for (D in 1:nrow(factors)) {
levels = unlist(strsplit(gsub("^.*\\(|)", "", factors[D,
2]), ","))
noEq = grep("=", levels, invert = TRUE)
levels[noEq] = paste(levels[noEq], "=never")
levels = matrix(unlist(strsplit(levels, "=")),
ncol = 2, byrow = TRUE)
x[, factors[D, 1]] = factor(x[, factors[D, 1]],
levels = as.numeric(levels[, 1]), labels = trimws(levels[,
2]))
}
 x$ID = factor(x$ID)
 # save(x, file = "../data/CF.Rdata")
 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")
# 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")
```

```r
 #CF_dat = x
mean_x = mean(x$FEV1)
sd_x = sd(x$FEV1)
# par(mfrow=c(1,2))
# hist(x$FEV1,
#      main="Evaluating Normality", freq=FALSE,
#      xlab="Lung Function Scores", ylab="Density")
# curve(dnorm(x, mean=mean_x, sd=sd_x),  col="red", add=TRUE)
# qqnorm(x$FEV1)
# qqline(x$FEV1)

#Original Model
x$ageC = x$AGE - 18
resS = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~1|ID, data = x)
model_random_intercept = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~1|ID, data = x)
# knitr::kable(summary(resS)$tTable, digits = 3, escape = FALSE, format = "latex", caption = '
#   kableExtra::kable_styling(latex_options = "hold_position")

#Random Slope Model
model_random_slope = lme(FEV1 ~ GENDER * F508 * ageC + PSEUDOA, random = ~ 1 + ageC | ID, data=
RS_table = summary(model_random_slope)$tTable
# knitr::kable(RS_table, digits=3, escape = FALSE, format = "latex", caption = 'Random Slope M
#   kableExtra::kable_styling(latex_options = "hold_position")

#Correlation Model
CF_time_model = lme(FEV1 ~ GENDER*F508*ageC + PSEUDOA, random = ~ 1| ID,
                    data=x, correlation = corExp(form=~ageC|ID, nugget=T))
corr_table = summary(CF_time_model)$tTable
# knitr::kable(corr_table, digits=3, escape = FALSE, format = "latex", caption = 'Serial Corre
#   kableExtra::kable_styling(latex_options = "hold_position")

# look at variogram for serial correlation

anova1 = anova(model_random_intercept,model_random_slope)

# knitr::kable(anova1[, c(0,3,4,9)], digits = 3,
#              caption = 'Random Slope Model is statistically significant (pvalue <0.05), whic
#   kableExtra::kable_styling(latex_options = "hold_position")
#
anova2 = anova(model_random_intercept, CF_time_model)
#
# knitr::kable(anova2[,c(0,3,4,9)], digits = 3,
#              caption = 'Serial Correlation Model is statistically significant (pvalue <0.05)
#   kableExtra::kable_styling(latex_options = "hold_position")

# Need to evaluate the Random Slope model vs Correlation model
# anova(model_random_slope, CF_time_model) # no p-value outputed
```

```r
count_tbl = sqldf("select
        ID,
        count(ID) as cnt_id
      from x
      group by 1")

min_pts = min(count_tbl$cnt_id)
max_pts = max(count_tbl$cnt_id)
par(mfrow=c(1,2))
a = plot(model_random_slope, resid(., type = "p") ~ fitted(.) , abline = 0)
b = plot(CF_time_model, resid(., type = "p") ~ fitted(.) , abline = 0)
# print(a, position = c(0, 0, 0.5, 1), more = TRUE)
# print(b, position = c(0.5, 0, 1, 1))

dataDir = "../data"
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)
forSurv = data.frame(time = (pmin(forInla$Age_first_tried_cigt_smkg, forInla$Age) - 4)/10,
                     event = forInla$Age_first_tried_cigt_smkg <= forInla$Age)
# we put this as event 2 because age 8 is the earliest that we record their smoking, so it mig
forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2

# Modifications to the prior according to the assignment
# inla.doc("pc.prec")
# pc.prec = penalized complexity on the precision of the random effects
#https://rdrr.io/github/mdsumner/INLA/man/pc-prec.html

# this is the state upper:
sigma1_prior = sqrt(2*log(5))
# this is the school upper: sqrt(2*log(1.5))
sigma2_prior = sqrt(2*log(1.5))

# Original Model
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(0.5,0.05)))) +
               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))

#Fitted model with priors
fitS3 = inla(y ~ RuralUrban + Sex * Race +
               f(school, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec", param = c(sigma2_prior,0.10)))) +
               f(state, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec", param = c(sigma1_prior,0.10)))),
            control.family = list(variant = 1,
                hyper = list(alpha = list(prior = "normal", param = c(1,(2/3)^(-2))))),
            data = forInla, family = "weibullsurv",
            control.compute=list(config = TRUE))

#Prob(1/sqrt(prec) > u) = alpha
# Graphing the histogram
# Seq = seq(0.4,1.5, len=1000)
# hist(forSurv$time, main='', xlab='Age', ylab='dens', freq=FALSE)
# kappa = fitS3$summary.hyper['alpha', 'mode']
# lambda = exp(-fitS3$summary.fixed['(Intercept)', 'mode'])
# lines(Seq, dweibull(Seq, shape=kappa, scale=lambda), col='blue')

#Graphing the hazard function

#plot random prior to posterior - slide 74 of survival
# par(mfrow=c(1,3))
# fitS3$priorPost = Pmisc::priorPost(fitS3)
# for (Dparam in fitS3$priorPost$parameters) {
#   do.call(matplot, fitS3$priorPost[[Dparam]]$matplot)
# }
# do.call(legend, fitS3$priorPost$legend)

#plot random prior to posterior - slide 74 of survival
# par(mfrow=c(1,3))
# fitS2$priorPost = Pmisc::priorPost(fitS2)
# for (Dparam in fitS2$priorPost$parameters) {
#   do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
# }
# do.call(legend, fitS2$priorPost$legend)
# sd of models of random effects - ph 50 of lecture 4
sdRes3 = Pmisc::priorPostSd(fitS3)
# knitr::kable(sdRes3$summary, digits = 3,
#              caption = 'Evaluation of Variation Between State and School')%>%
# kableExtra::kable_styling(latex_options = "hold_position")
```

```r
# Table of results of model
resTable3 = rbind(exp(fitS3$summary.fixed[,c(4,3,5)]), fitS3$summary.hyper[,c(4,3,5)])
colnames(resTable3) = c('Estimate', 'Lower 2.5%', 'Upper 97.5%')
# knitr::kable(resTable3[,], digits=3,
#              caption = 'Summary of Model Parameters')%>%
#   kableExtra::kable_styling(latex_options = "hold_position")

xSeq_3 = seq(12, 80, len=1000)
xSeq_3f = (xSeq_3 - 4)

forInla2 = smoke[, c("Age", "Age_first_tried_cigt_smkg",
                     "Sex", "Race", "state", "school", "RuralUrban")]
forInla2 <- sqldf("select *
                  from forInla2 where Sex='M'
                  and Race = 'white'
                  ")
forInla2 = na.omit(forInla2)
forInla2 = as.list(forInla2)
forSurv = data.frame(time = (pmin(forInla2$Age_first_tried_cigt_smkg, forInla2$Age) - 4)/10,
                     event = forInla2$Age_first_tried_cigt_smkg <= forInla2$Age)
forSurv[forInla2$Age_first_tried_cigt_smkg == 8, "event"] = 2

newCov = model.matrix(~RuralUrban, data.frame(RuralUrban=unique(forInla2$RuralUrban)))
rownames(newCov) = as.character(unique(forInla2$RuralUrban))

densHaz = Pmisc::sampleDensHaz(fit=fitS3, x=xSeq_3f, n=20, covariates = newCov, scale=10)
Scol = mapmisc::col2html(2:3, 0.2)
names(Scol) = c('Urban','Rural')

# plot(NA, xlim = c(12, 60), ylim = range(densHaz[, names(Scol), "dens", ]), xlab = "age", yla
# for (D in names(Scol)) { matlines(xSeq_3, densHaz[, D, "dens", ], type = "l", lty = 1, col =
# legend("topright", lty = 1, col = Scol, legend = names(Scol), bty = "n")
```