STA303 Midterm2 复习讲义

陈漂亮　　Savvy 2020

目录

- For dear readers, you are welcomed to find my source file of this document at my personal website.

0. 考试纪要

考试时间：占比 20%.

目前学过的内容（不一定都考的 hhh）：Week 4-5：

1. linear mixed model(LMM 概论)
2. LMM 的 loglike function 及 profile loglike function；
3. Cholesky Decomposition；
4. Woodbury Matrix Identity;
5. Restricted Maximum Likelihood (REML)
6. Fixed effect vs Random effect Model
7. Random Slope Model, Contrast model, multi level LMM
8. GLMM & ZIP

Week 6:

1. Case-Control Study

R code 的分析（重中之重）

如何写一篇 statistica report

1. 我们目前学过的 Model 一览

1.1 Linear Mixed Model (LMM)

$$
\begin{aligned}
& Y_{ij}|U_i \overset{iid}{\sim} N(\mu_{ij}, \tau^2) \\
& \mu_{ij} = X_{ij}\beta + U_i \\
& [U_1, \cdots, U_M]^T \sim MVN(0, \Sigma)(\text{where } \Sigma = \sigma I) \\
\iff & U_i \sim N(0, \sigma^2), i = 1, ..., M \\
\text{Where, } & \bullet Y_{ij} \text{ is the jth obs in the ith group}, j = 1, ..., J_i \\
& \bullet U_i \text{ is the individual's random effect}, i = 1, ..., M \\
& \bullet X_{ij}\beta \text{ is the fixed effect}
\end{aligned}
$$

1.1.1 Fixed Effect Model

$$
\begin{aligned}
\text{Fixed: } & Y_{ij} = X_{ij}\beta + \theta_i + Z_{ij}, Z_{ij} \sim N(0, \tau^2) \\
\Longrightarrow & \hat{\theta} = \arg\max_{\theta_i}(P(Y_{ij}; \theta_i)) \text{ , which can be considered as intercept} \\
\text{Random: } & Y_{ij} = X_{ij}\beta + U_i + Z_{ij}, U_i \sim N(0, \sigma^2), Z_{ij} \sim N(0, \tau^2) \\
\Longrightarrow & \hat{U}_i = E(U_i|\mathbf{Y})
\end{aligned}
$$

1.1.2 Constrast Model

$$
E(Y_{ij}) = \begin{cases} \theta_j, & diet_i = \text{barley} \\ \theta_j + \alpha_1 + \beta_1 t_j + (\beta_1' t_j^2), & diet_i = \text{lupins} \\ \theta_j + \alpha_2 + \beta_2 t_j + (\beta_2' t_j^2), & diet_i = \text{mixed} \end{cases}
$$

$$
\begin{aligned}
\iff E(Y_{ij}) = \mu_{ij} = \theta_j + I_{lupins} \times \alpha_1 + I_{mixed} \times \alpha_2 \\
+ (I_{lupins} \times \beta_1 + I_{mixed} \times \beta_2)t_j + (I_{lupins} \times \beta_1 + I_{mixed} \times \beta_2)t_j^2
\end{aligned}
$$

1.1.3 Multi-level models

$$
\begin{aligned}
& Y_{ijkm}|B, C, D \sim N(\mu_{ijkm}, \tau^2) \\
& \mu_{ijkm} = X_{ijkm}\beta + B_i + C_{ij} + D_{ijk} \\
\text{where, } & B_i \sim N(0, \sigma_B^2) \\
& B_i \sim N(0, \sigma_B^2) \\
& C_{ij} \sim N(0, \sigma_C^2) \\
& D_{ijk} \sim N(0, \sigma_D^2)
\end{aligned}
$$

1.1.4 Random Slope Model

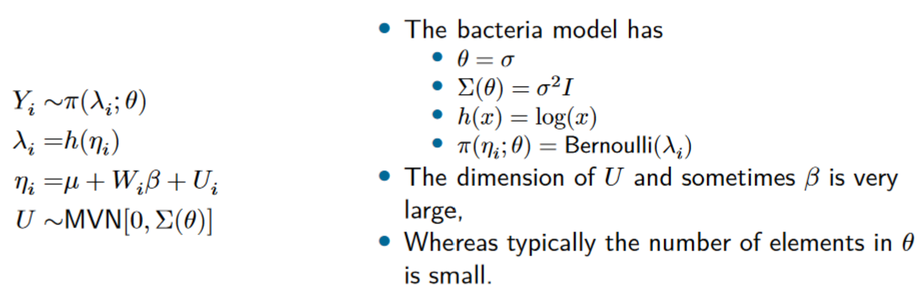
$$
\begin{aligned}
& Y_{ij}|\mathbf{U} \sim N(\mu_{ij}, \tau^2) \\
\text{where, } & \mu_{ij} = \mathbf{X}_{ij}\beta + U_{i1} + U_{i2}W_{ij}; \\
& \begin{pmatrix} U_{i1} & \text{(random intercept)} \\ U_{i2} & \text{(random slope)} \end{pmatrix} \sim MVN(0, \Gamma) \\
& X_{ijp}, W_{ij} \text{ are all covariates, but the coef on } W_{ij} \text{ is different for each subject}
\end{aligned}
$$

1.2 Generalized Linear Mixed Model (GLMM)

$$
\begin{aligned}
& Y_{ij}|U \overset{\perp}{\sim} G(\mu_{ij}, \theta) \text{ , for some distribution G} \\
& h(\mu_{ij}) = X_{ij}\beta + U_i, \text{ where h() is the link function} \\
& U \sim MVN(0, \Sigma)
\end{aligned}
$$

Patrick's definition:

```
knitr::include_graphics("1.png")
```



1.2.1 Binomial GLMM

$$
\begin{aligned}
& Y_{it} \overset{\perp}{\sim} Bernoulli(\rho_{it}) \\
& logit(\rho_{it}) = \mu + X_{it}\beta + U_i \\
& U_i \overset{\perp}{\sim} N(0, \sigma^2)
\end{aligned}
$$

where, • $Y_{ij}$is 1 if the ith individual is infected at time t

• X_{it} has indicator variable for week and treatment type

• U_i is the individual level random effect

$U_i > 0$ if the ith individual is more likely to be infected than the average, allowng for within-ind dependence

• σ is the extra-Bernoulli variation, or overdispersion

• ρ_{it} is the prob of the ith individual being infected at time t

1.2.2 Gamma GLMM

$$
\begin{aligned}
&Y_{ij} \sim Gamma(\frac{\mu_{ij}}{v}, v) \\
&\log(\mu_{ij}) = \beta_0 + t_{ij}\beta_1 + U_i \\
&U_i \sim N(0, \sigma^2)
\end{aligned}
$$

where,
- Y_{ij} is the weight of pig i at time t_{ij}
- $\exp(\beta_0)$ is the pop average weight at birth
- $\exp(\beta_1)$ is the average porportion weight gain, per year
- U_i is the pig i's weight deviation from the pop average

1.2.3 Zero Inflated Poisson

$$
\begin{aligned}
&Y_i|U \sim ZIP(O_i\lambda_i, \rho) \\
&\log(\lambda_i) = X_i\beta + U_i \\
&U_i \sim N(0, \sigma^2)
\end{aligned}
$$

where,
- $Y_i = \begin{cases} 1, \text{ with prob } 1-\theta \\ 0, \text{ with prob } \theta \end{cases}$
- O_i is the offset term for the ith level
- X_i is the covariate of the ith level
- U_i is the i-th level random effect
- σ is the extra-Poisson variation, or overdispersion
- ρ is the porportion of couples which are infertile

1.3 Case-Control Study

Want: $P(Y_i|X_i), Y_i \sim Binom(N_i, \mu_i)$

Have: $P(Y_i|X_i, Z_i)$

Where,
- $Y_i = 1$ is the case; $= 0$ is the control;
- X_i are the covariates, e.g. conditions/enviornment when case/control occurs;
- Z_i is "in-the-study" indicator, $= 1$ if the data is observed

Assumption:

$$P(Z_i|Y_i, X_i) = P(Z_i|Y_i)$$

i.e., Inclusion in the study does not depend on covariate.

Critiques of this assumption:

(a) Control are hard to recruit: man tend not to report light accident than woman.

- Underestiamte the covariate/case effect, e.g. sex/fatal-accident effect…

(b) Controls are easy to recruit: healthy smokers tend to worry about cancer if have a family histroy of lung cancer.

- Overestimate the covariate/case effect, e.g. smoking/cancer effect…

2. 考题类型

2.1 Write down the model: see above

2.2 How good/bad is this model: Model assessment

- Anova: normality, independence, homoskedasticity

- LMM vs OLM: whether we should include the random effect

- GLMM: model adequacy (whether the data follows the distribution the model assumes); should/Not include random effect

- Other: Histogram, or see the nature of the question

2.3 Tests

- Likelihood Ratio Test;

- What to look at for a test? p-value

2.4 Model result interpretation: read R code and R output

2.4.1 LMM

```
#'install.packages("Pmisc",
#     'repos='http://r-forge.r-project.org')
#+
cUrl =  'https://www.fueleconomy.gov/feg/epadata/vehicles.csv.zip'
cFile = file.path(tempdir(), basename(cUrl))
download.file(cUrl, cFile)
cFile2 = unzip(cFile, exdir=tempdir())
x = read.table(cFile2, sep=',', header=TRUE, stringsAsFactors=FALSE)
#'
```

```r
#' https://www.fueleconomy.gov/feg/ws/index.shtml#vehicle
xSub = x[grep("Electricity|CNG",
              x$fuelType, invert=TRUE), ]
#'
#+  makeFac: factorize the names of the auto makers
#+  cylFac: factorize the number of cylinder and change the baseline to "4"
xSub$decade = (xSub$year - 2000)/10
makeTable = sort(table(xSub$make), decreasing=TRUE)
xSub$makeFac = factor(xSub$make, levels=names(makeTable))
xSub$cylFac = relevel(factor(xSub$cylinders), '4')
# levels(xSub$cylFac)
xSub = xSub[!is.na(xSub$cylFac), ]
xSub$transmission = factor(
  grepl("Manual", xSub$trany), levels=c(FALSE,TRUE),
    labels = c('Automatic', 'Manual'))#'
#'
```

- Fit the LMM model, using makeFac as the random effect. There are 2 ways to fit the LMM models, and the way they restore the data is quite different.

```r
#+ lmm
# comb08: combined MPG, measures fuel efficiency
myFitMake = nlme::lme(comb08 ~ cylFac +
                  decade + transmission,
                  random = ~1|makeFac,
                data=xSub)
# summary(myFitMake)$tTable[,1:2]
Pmisc::lmeTable(myFitMake)
```

```
##                       MLE  Std.Error    DF     t-value       p-value
## (Intercept)        23.3363393 0.19089242 41547  122.248641  0.000000e+00
## cylFac2            -5.8542655 0.43748620 41547  -13.381600  9.409591e-41
## cylFac3             9.7958373 0.19525174 41547   50.170295  0.000000e+00
## cylFac5            -3.7580231 0.12857605 41547  -29.228017 6.569783e-186
## cylFac6            -5.6040722 0.03920337 41547 -142.948745  0.000000e+00
## cylFac8            -8.5333559 0.04867933 41547 -175.297302  0.000000e+00
## cylFac10          -10.2526438 0.26084048 41547  -39.306183  0.000000e+00
## cylFac12          -10.3261759 0.15494310 41547  -66.644955  0.000000e+00
## cylFac16          -14.6612041 2.04593093   129   -7.166031  5.294830e-11
## decade              1.2036653 0.01537899 41547   78.266884  0.000000e+00
```

```
## transmissionManual   0.5003242 0.03537212 41547   14.144591  2.570305e-45
## $\\sigma$            1.8184385         NA    NA          NA            NA
## $\\tau$              3.0440381         NA    NA          NA            NA
```

```
nlme::intervals(myFitMake)$sigma
```

```
##    lower     est.    upper
## 3.023411 3.044038 3.064805
## attr(,"label")
## [1] "Within-group standard error:"
```

```
nlme::intervals(myFitMake)$reStruct$makeFac
```

```
##                    lower     est.    upper
## sd((Intercept)) 1.529923 1.818438 2.161362
```

```
#'
#+ lme4
myFitMake2 = lme4::lmer(comb08 ~ cylFac +
                  decade + transmission +
                  (1|makeFac),
                data=xSub)
summary(myFitMake2)$coef
```

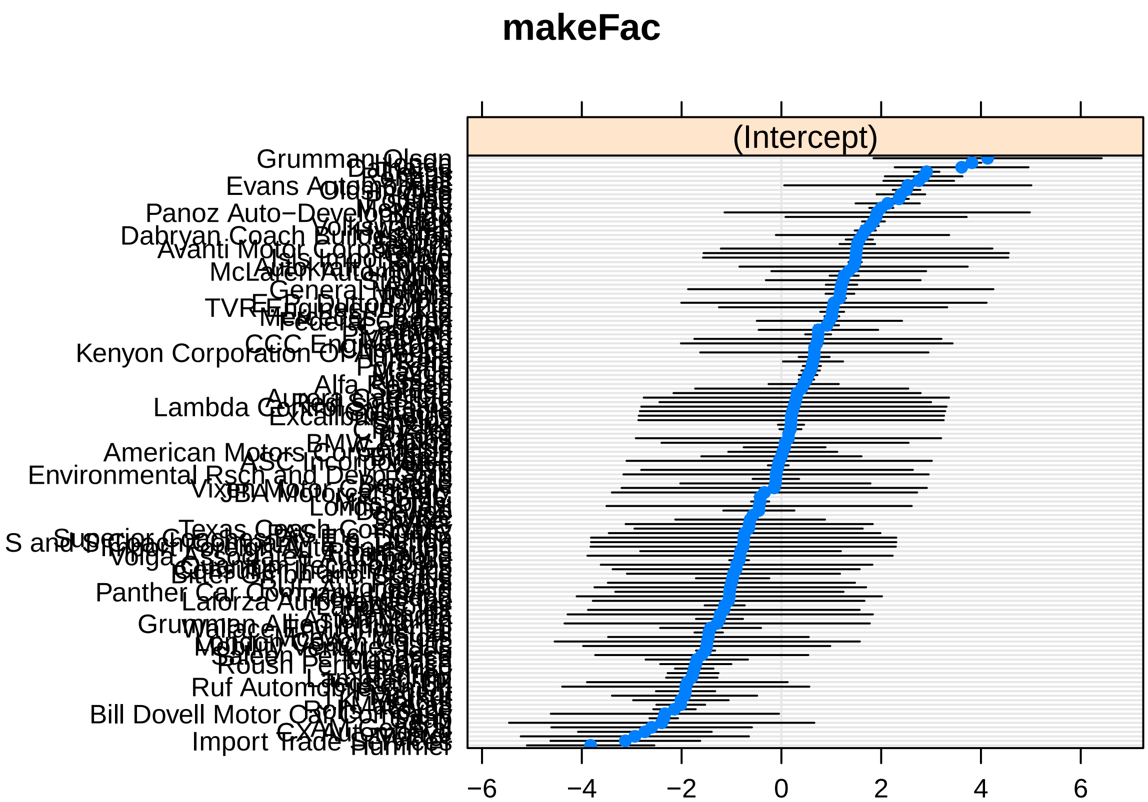
```
##                       Estimate Std. Error     t value
## (Intercept)         23.3363393 0.19089242  122.248643
## cylFac2             -5.8542655 0.43748620  -13.381600
## cylFac3              9.7958373 0.19525174   50.170295
## cylFac5             -3.7580231 0.12857605  -29.228017
## cylFac6             -5.6040722 0.03920337 -142.948745
## cylFac8             -8.5333559 0.04867933 -175.297302
## cylFac10           -10.2526438 0.26084048  -39.306183
## cylFac12           -10.3261760 0.15494310  -66.644955
## cylFac16           -14.6612041 2.04593091   -7.166031
## decade               1.2036653 0.01537899   78.266884
## transmissionManual   0.5003242 0.03537212   14.144591
```

```
summary(myFitMake2)$varcor
```

```
##  Groups   Name        Std.Dev.
##  makeFac  (Intercept) 1.8184
##  Residual             3.0440
```

```r
# plot the random effect for the lme4::lmer model, using lme4::ranef
myFitRandom = lme4::ranef(myFitMake2, condVar=TRUE)
lattice::dotplot(myFitRandom)
```

```
## $makeFac
```



- A fancier random effect plot: what can you tell about the random effect of auto-maker on the combined MPG from the 2 plots?

```r
x = data.frame(
    make = rownames(myFitRandom$makeFac),
    est = myFitRandom$makeFac[[1]],
    se = drop(attributes(myFitRandom$makeFac)$postVar),
    stringsAsFactors = FALSE
    )
x$lower = x$est - 2*x$se
x$upper = x$est + 2*x$se
x = x[x$se < 2, ]
x = x[order(x$est), ]
```

```r
x$index = rank(x$est)
x$accurate = rank(x$se) < 40

x$col= rep_len(RColorBrewer::brewer.pal(8, 'Set2'), nrow(x))
x$colTrans = paste0(x$col, '40')
x$colLine = x$col
x[!x$accurate,'colLine'] = x[!x$accurate,'colTrans']

x$cex = -log(x$se)
x$cex = x$cex - min(x$cex)
x$cex = 3*x$cex / max(x$cex)



x$textpos = rep_len(c(4,2), nrow(x))
x[!x$accurate & x$est > 0, 'textpos'] = 4
x[!x$accurate & x$est < 0, 'textpos'] = 2

x$textloc = x$est



x$textCex = c(0.5, 0.9)[1+x$accurate]



par(mar=c(4,0,0,0), bty='n')
plot(x$est, x$index, yaxt='n', xlim = range(x$est),
    #xlim = range(x[,c('lower','upper')]),
    xlab='mpg', ylab='', pch=15, col=x$colTrans , cex=x$cex)

x[!x$accurate & x$est > 0, 'textloc'] = par('usr')[1]
x[!x$accurate & x$est < 0, 'textloc'] = par('usr')[2]

abline(v=0, col='grey')
segments(x$lower, x$index, x$upper, x$index, pch=15, col=x$colLine)
text(
    x$textloc,
    x$index, x$make,
    pos = x$textpos,
    col=x$col,
    cex=x$textCex, offset=1)
```

图 1: Fuel efficiency of auto manufacturers adjusted for engine type

```
#'
#'
```

2.4.2 GLMM

2.4.2.1 Gamma

- Fit the GLMM models, in 2 ways: `lme4::glmer`, or `glmmTMB`.

```
install.packages("Pmisc", repos='http://r-forge.r-project.org')
```

```
## package 'Pmisc' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\maxch\AppData\Local\Temp\RtmpSOxaKu\downloaded_packages
```

```
#+
cUrl =  'https://www.fueleconomy.gov/feg/epadata/vehicles.csv.zip'
cFile = file.path(tempdir(), basename(cUrl))
download.file(cUrl, cFile)
cFile2 = unzip(cFile, exdir=tempdir())
x = read.table(cFile2, sep=',', header=TRUE, stringsAsFactors=FALSE)
#'
#' https://www.fueleconomy.gov/feg/ws/index.shtml#vehicle
#+
xSub = x[grep("Electricity|CNG",
              x$fuelType, invert=TRUE), ]
#'
#+  makeFac
xSub$decade = (xSub$year - 2000)/10
makeTable = sort(table(xSub$make), decreasing=TRUE)
xSub$makeFac = factor(xSub$make, levels=names(makeTable))
xSub$cylFac = relevel(factor(xSub$cylinders), '4')
levels(xSub$cylFac)
```

```
## [1] "4"  "2"  "3"  "5"  "6"  "8"  "10" "12" "16"
```

```
xSub = xSub[!is.na(xSub$cylFac), ]
xSub$transmission = factor(
  grepl("Manual", xSub$trany), levels=c(FALSE,TRUE),
```

```
    labels = c('Automatic', 'Manual'))#'
#'

#' Gaussian
#+ lme4
myFitMakeL = lme4::lmer(comb08 ~ cylFac +
                  decade + transmission +
                  (1|makeFac),
                data=xSub)
summary(myFitMakeL)$coef
```

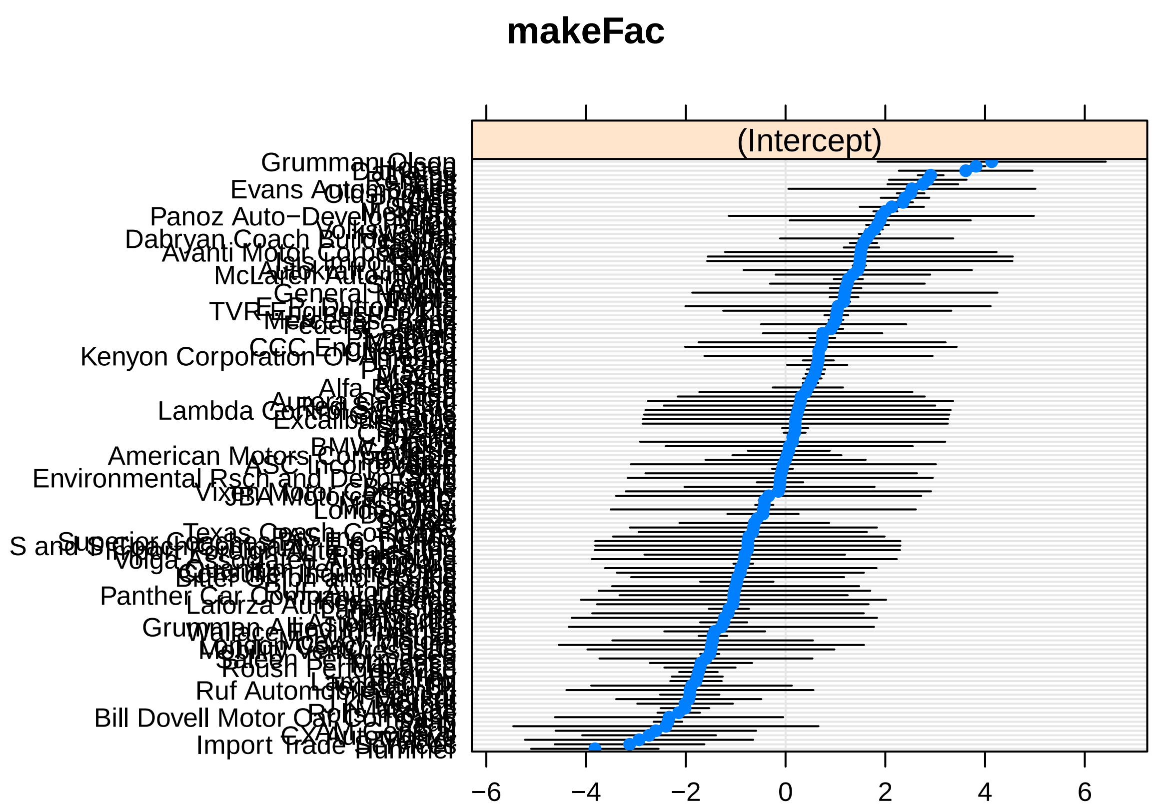
```
##                       Estimate Std. Error     t value
## (Intercept)         23.3363393 0.19089242  122.248643
## cylFac2             -5.8542655 0.43748620  -13.381600
## cylFac3              9.7958373 0.19525174   50.170295
## cylFac5             -3.7580231 0.12857605  -29.228017
## cylFac6             -5.6040722 0.03920337 -142.948745
## cylFac8             -8.5333559 0.04867933 -175.297302
## cylFac10           -10.2526438 0.26084048  -39.306183
## cylFac12           -10.3261760 0.15494310  -66.644955
## cylFac16           -14.6612041 2.04593091   -7.166031
## decade               1.2036653 0.01537899   78.266884
## transmissionManual   0.5003242 0.03537212   14.144591
```

```
summary(myFitMakeL)$varcor
```

```
##  Groups   Name        Std.Dev.
##  makeFac  (Intercept) 1.8184
##  Residual             3.0440
```

```
myFitRandom = lme4::ranef(myFitMakeL, condVar=TRUE)
lattice::dotplot(myFitRandom)
```

```
## $makeFac
```




```r
#'

#'
#+ glmer
myFitMakeG = lme4::glmer(comb08 ~ cylFac+transmission +
                    decade + (1|makeFac),
                  family=Gamma(link=log), data=xSub[xSub$year < 1990, ])
# myFitMakeG$table = Pmisc::coefTable(myFitMakeG)

summary(myFitMakeG)$coef
```

```
##                        Estimate  Std. Error    t value     Pr(>|z|)
## (Intercept)          2.97165005 0.027586421 107.721478 0.000000e+00
## cylFac2             -0.29215655 0.034658380  -8.429608 3.468260e-17
## cylFac3              0.39827430 0.020500037  19.427980 4.476373e-84
## cylFac5             -0.15961619 0.020017246  -7.973934 1.537019e-15
## cylFac6             -0.27817975 0.004562385 -60.972442 0.000000e+00
## cylFac8             -0.44116104 0.005071318 -86.991399 0.000000e+00
## cylFac12            -0.56073390 0.042503870 -13.192538 9.686864e-40
```

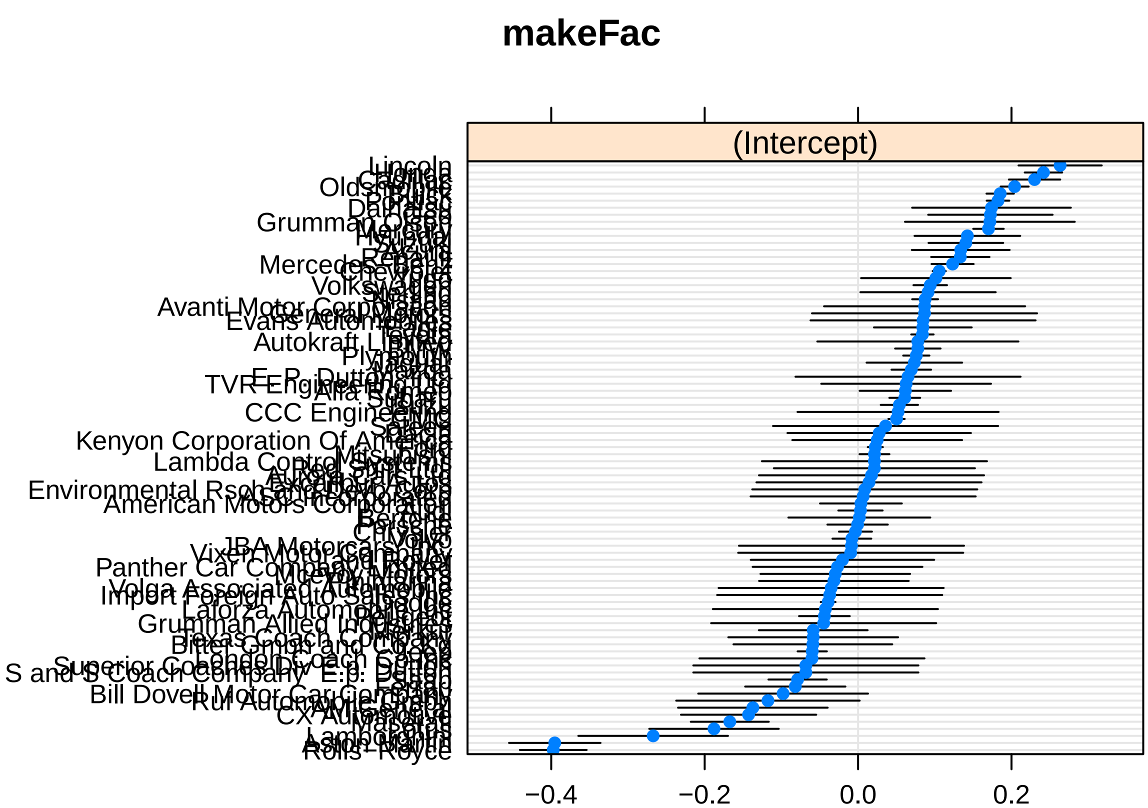
```
## transmissionManual  0.06051229 0.003426021  17.662557 8.146435e-70
## decade             -0.04083182 0.009327289  -4.377673 1.199531e-05
```

```r
lme4::VarCorr(myFitMakeG)
```

```
##  Groups   Name        Std.Dev.
##  makeFac  (Intercept) 0.08659
##  Residual             0.14899
```

```r
lattice::dotplot(lme4::ranef(myFitMakeG, condVar=TRUE))
```

```
## $makeFac
```

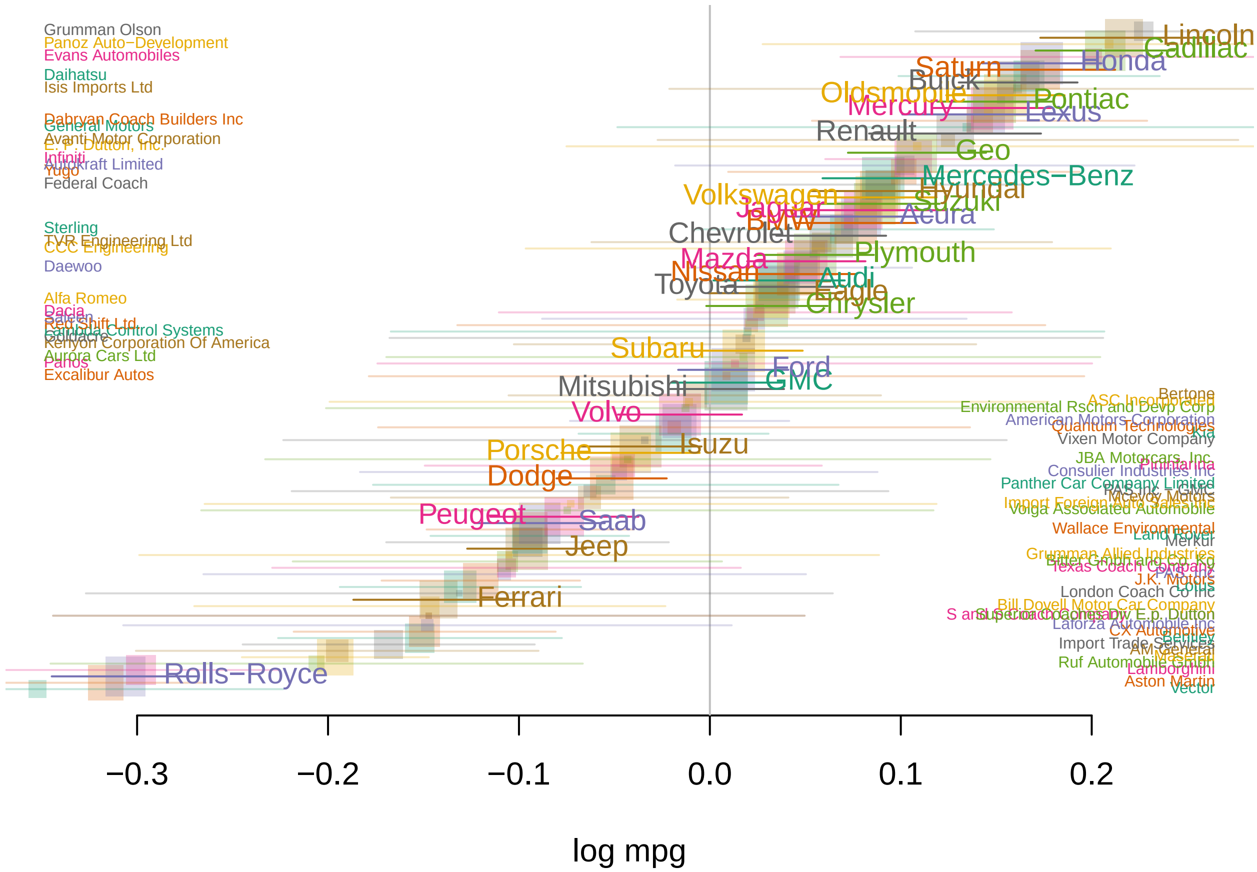



```r
#'



#'
#+ glmmTMB

myFitMakeTmb = glmmTMB::glmmTMB(comb08 ~ cylFac+transmission +
```

```
                     decade + (1|makeFac),
                   family=Gamma(link=log), data=xSub[xSub$year < 2000, ])
# myFitMakeTmb$table = Pmisc::coefTable(myFitMakeTmb)

knitr::kable(myFitMakeTmb$table, digits=2)
```

|| || || ||

```
Pmisc::ranefPlot(x=myFitMakeTmb, xlab= 'log mpg')
```




```
#'
```

2.4.2.2 ZIP

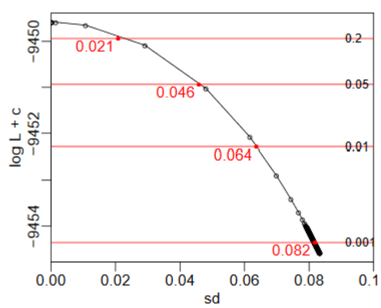
- What you predicted is number of event happen in a given time interval (e.g. annual birth rate = number of children born to a mother in one year);
- Why sd is from 0 to infinity? Uncorrected LR test failed here, as the the distribution assumption is false.

```r
fResZI = glmmTMB::glmmTMB(children~ offset(logYears)+ethnicity
                    +literacy*residence+ageMarried+(1|subject),
                    ziformula =~1,
                    data=fiji,family=poisson)
toPrint =Pmisc::coefTable(+fResZI)$table
```

	est	2.5 %	97.5 %
baseline			
fijian:yes:suva:0to15	0.24	0.23	0.26
literacy			
no	1.05	0.96	1.15
residence			
otherUrban	1.11	1.04	1.17
rural	1.20	1.14	1.26
literacy:residence			
no:otherUrban	0.98	0.87	1.10
no:rural	0.93	0.85	1.03
sd			
subject	0.00	0.00	Inf
zero inf			
prob	0.02	0.02	0.03

	est	2.5 %	97.5 %
ethnicity			
indian	0.98	0.94	1.02
european	0.79	0.57	1.10
partEuropean	0.99	0.86	1.13
pacificIslander	1.13	1.02	1.27
routman	0.93	0.72	1.21
chinese	0.76	0.58	0.99
other	1.86	1.10	3.14
ageMarried			
15to18	1.11	1.06	1.16
18to20	1.16	1.11	1.22
20to22	1.14	1.08	1.21
22to25	1.14	1.07	1.22
25to30	1.20	1.08	1.33
30toInf	1.30	1.07	1.59

- right numbers are significance levels
- Red lines are corresponding quantiles of the LR distribution
- 99% CI for σ is $(0, 0.062)$



2.4.2.3 Binomial + Multi-level random effect

```r
school$mathWrong =40-school$math
schoolGlmm =glmmTMB::glmmTMB(cbind(math, mathWrong)~gender
                        +socialClass+(1|school/class/student),
                        data =school, family =binomial(link ="logit"))
```

- Describe the model used here.

```r
Pmisc::mdTable(Pmisc::coefTable(+schoolGlmm)$table,guessGroup=F,dec=2)
```

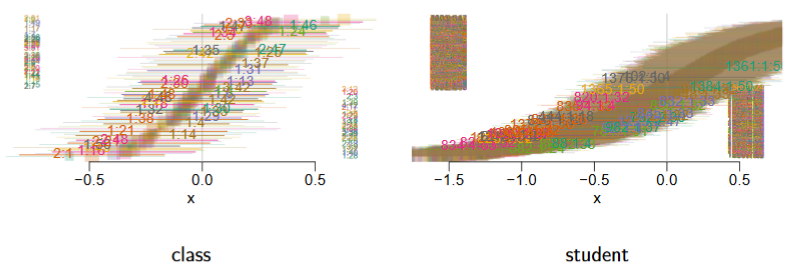
```r
knitr::include_graphics("3.png")
```

	est	2.5 %	97.5 %
ref prob			
f:l	0.76	0.71	0.81
gender			
m	0.99	0.90	1.09
socialClass			
II	1.00	0.74	1.36
IIIn	0.75	0.55	1.04
IIIm	0.63	0.47	0.84
IV	0.66	0.48	0.91
V	0.51	0.37	0.71
longUnemp	0.58	0.41	0.82
currUnemp	0.52	0.33	0.82
absent	0.59	0.44	0.80
sd			
student:(class:school)	0.74	0.71	0.78
class:school	0.28	0.22	0.37
school	0.00	0.00	Inf

- Compare individual/student-level effect and class-level effect

```
Pmisc::ranefPlot(schoolGlmm,grpvar ="class:school")
Pmisc::ranefPlot(schoolGlmm,grpvar ="student:(class:school)")
```

```
knitr::include_graphics("4.png")
```



- What happened to the school-level effect?
  Effective sample size is too small for the school level variation, because we have too few schools in our data set. E.g. suppose we have 2 schools, each school has 15 classes and each class has 60 students. Then the effective sample size for individual/student-level effect is $2 \times 15 \times 60$, for class-level effect is 30, and for school-level effect is only 2. The smaller the effective sample size, the larger the SD for the estiamtor $\sigma_i's$, and thus the wider the CI.

2.4.3 Case-Control Study

GLM: pedestrian

```
#+ pedestrianData, eval=FALSE
pedestrainFile = Pmisc::downloadIfOld(
  'http://pbrown.ca/teaching/303/data/pedestrians.rds')
pedestrians = readRDS(pedestrainFile)
```

```r
pedestrians = pedestrians[!is.na(pedestrians$time), ]
pedestrians$y = pedestrians$Casualty_Severity == 'Fatal'
#'

#'
#'
#+ showPedestrians
dim(pedestrians)
```

```
## [1] 1159371       7
```

```r
pedestrians[1:3,]
```

```
##                   time     age  sex Casualty_Severity
## 54 1979-01-01 22:40:00 26 - 35 Male            Slight
## 65 1979-01-02 10:40:00 26 - 35 Male            Slight
## 79 1979-01-02 14:25:00 46 - 55 Male            Slight
##         Light_Conditions    Weather_Conditions     y
## 54 Darkness - lights lit Snowing no high winds FALSE
## 65              Daylight Raining no high winds FALSE
## 79              Daylight Raining no high winds FALSE
```

```r
table(pedestrians$Casualty_Severity, pedestrians$sex)
```

```
##
##             Male Female
##   Slight 637919 481811
##   Fatal   24429  15212
```

```r
range(pedestrians$time)
```

```
## [1] "1979-01-01 01:00:00 EST" "2015-12-31 23:35:00 EST"
```

```r
table(pedestrians$age, pedestrians$Casualty_Severity)
```

```
##
##            Slight  Fatal
##   26 - 35 118457   3083
##   0 - 5     84233   1608
##   6 - 10   168054   1976
##   11 - 15 197505   2219
```

```
##   16 - 20 128073   2551
##   21 - 25  86776   2104
##   36 - 45  86247   2989
##   46 - 55  69942   3138
##   56 - 65  62342   4011
##   66 - 75  59699   6235
##   Over 75  58402   9727
```

```
#'
#'
#+ pedGlm
theGlm = glm(y ~ sex + age + Light_Conditions + Weather_Conditions, data=pedestrians,
  family=binomial(link="logit") )
theTable = as.data.frame(summary(theGlm)$coef)
theTable
```

```
##                                              Estimate Std. Error
## (Intercept)                               -4.17701178 0.02048265
## sexFemale                                 -0.27468850 0.01113694
## age0 - 5                                   0.18571678 0.03184733
## age6 - 10                                 -0.35712560 0.02968724
## age11 - 15                                -0.50420038 0.02853721
## age16 - 20                                -0.33807211 0.02748975
## age21 - 25                                -0.15867913 0.02907749
## age36 - 45                                 0.32447065 0.02656711
## age46 - 55                                 0.65995081 0.02636673
## age56 - 65                                 1.13797354 0.02509062
## age66 - 75                                 1.75962977 0.02338872
## ageOver 75                                 2.32777540 0.02231760
## Light_ConditionsDarkness - lights lit      0.99467431 0.01224661
## Light_ConditionsDarkness - lights unlit    1.17557865 0.05244657
## Light_ConditionsDarkness - no lighting     2.76545143 0.02106167
## Light_ConditionsDarkness - lighting unknown  0.25940010 0.06847541
## Weather_ConditionsRaining no high winds   -0.21426159 0.01653660
## Weather_ConditionsSnowing no high winds   -0.75145427 0.09235823
## Weather_ConditionsFine + high winds        0.17495735 0.03664924
## Weather_ConditionsRaining + high winds    -0.06591513 0.03999709
## Weather_ConditionsSnowing + high winds    -0.54970673 0.17213835
## Weather_ConditionsFog or mist              0.06850742 0.06926016
##                                                z value      Pr(>|z|)
```

```
## (Intercept)                                      -203.9292454  0.000000e+00
## sexFemale                                         -24.6646232 2.564554e-134
## age0 - 5                                            5.8314718  5.494058e-09
## age6 - 10                                         -12.0296000  2.483611e-33
## age11 - 15                                        -17.6681741  7.374545e-70
## age16 - 20                                        -12.2981159  9.271148e-35
## age21 - 25                                         -5.4571129  4.839387e-08
## age36 - 45                                         12.2132473  2.641326e-34
## age46 - 55                                         25.0296758 2.906540e-138
## age56 - 65                                         45.3545452  0.000000e+00
## age66 - 75                                         75.2341309  0.000000e+00
## ageOver 75                                        104.3022110  0.000000e+00
## Light_ConditionsDarkness - lights lit              81.2203565  0.000000e+00
## Light_ConditionsDarkness - lights unlit            22.4147881 2.823821e-111
## Light_ConditionsDarkness - no lighting            131.3025425  0.000000e+00
## Light_ConditionsDarkness - lighting unknown         3.7882226  1.517289e-04
## Weather_ConditionsRaining no high winds           -12.9568106  2.150027e-38
## Weather_ConditionsSnowing no high winds            -8.1362997  4.075424e-16
## Weather_ConditionsFine + high winds                 4.7738334  1.807520e-06
## Weather_ConditionsRaining + high winds             -1.6479982  9.935303e-02
## Weather_ConditionsSnowing + high winds             -3.1934007  1.406077e-03
## Weather_ConditionsFog or mist                       0.9891318  3.225986e-01
```

```r
#'
#'
#+ ci95
theTable$low = theTable$Estimate - 2*theTable$'Std. Error'
theTable$high = theTable$Estimate + 2*theTable$'Std. Error'
exp(theTable[,c('Estimate','low','high')])
```

```
##                                                   Estimate        low
## (Intercept)                                     0.01534429 0.01472841
## sexFemale                                       0.75980876 0.74307196
## age0 - 5                                        1.20408119 1.12977910
## age6 - 10                                       0.69968461 0.65935047
## age11 - 15                                      0.60398834 0.57048135
## age16 - 20                                      0.71314386 0.67499391
## age21 - 25                                      0.85327010 0.80506351
## age36 - 45                                      1.38329820 1.31171630
## age46 - 55                                      1.93469717 1.83531723
```

```
## age56 - 65                                    3.12043850  2.96771502
## age66 - 75                                    5.81028586  5.54475446
## ageOver 75                                   10.25510260  9.80742929
## Light_ConditionsDarkness - lights lit         2.70384358  2.63842219
## Light_ConditionsDarkness - lights unlit       3.24001725  2.91737875
## Light_ConditionsDarkness - no lighting       15.88620985 15.23092771
## Light_ConditionsDarkness - lighting unknown   1.29615229  1.13026179
## Weather_ConditionsRaining no high winds       0.80713722  0.78087922
## Weather_ConditionsSnowing no high winds       0.47168010  0.39212652
## Weather_ConditionsFine + high winds           1.19119541  1.10700578
## Weather_ConditionsRaining + high winds        0.93621032  0.86423608
## Weather_ConditionsSnowing + high winds        0.57711904  0.40902319
## Weather_ConditionsFog or mist                 1.07090858  0.93238180
##                                                     high
## (Intercept)                                   0.01598593
## sexFemale                                     0.77692255
## age0 - 5                                      1.28326990
## age6 - 10                                     0.74248611
## age11 - 15                                    0.63946335
## age16 - 20                                    0.75345000
## age21 - 25                                    0.90436327
## age36 - 45                                    1.45878641
## age46 - 55                                    2.03945840
## age56 - 65                                    3.28102139
## age66 - 75                                    6.08853324
## ageOver 75                                   10.72321056
## Light_ConditionsDarkness - lights lit         2.77088714
## Light_ConditionsDarkness - lights unlit       3.59833697
## Light_ConditionsDarkness - no lighting       16.56968427
## Light_ConditionsDarkness - lighting unknown   1.48639083
## Weather_ConditionsRaining no high winds       0.83427818
## Weather_ConditionsSnowing no high winds       0.56737331
## Weather_ConditionsFine + high winds           1.28178780
## Weather_ConditionsRaining + high winds        1.01417862
## Weather_ConditionsSnowing + high winds        0.81429707
## Weather_ConditionsFog or mist                 1.23001669
```

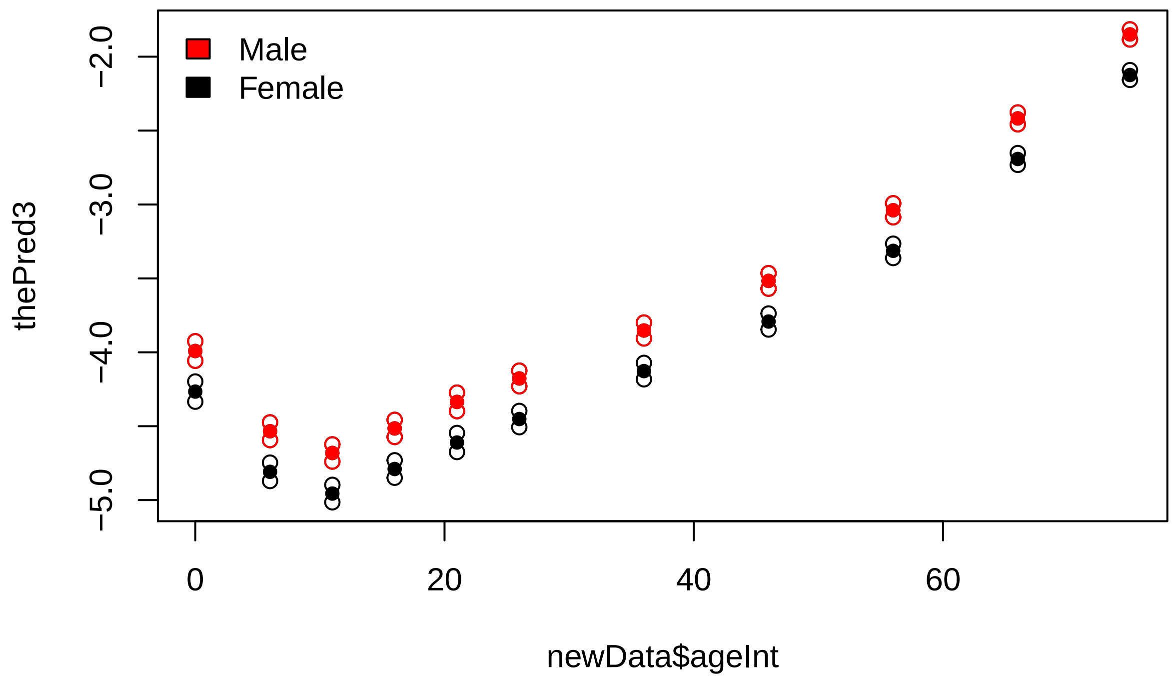
```r
#'
#'
#+ newdata
```

```
newData = expand.grid(
    age = levels(pedestrians$age),
    sex = c('Male', 'Female'),
    Light_Conditions = levels(pedestrians$Light_Conditions)[1],
    Weather_Conditions = levels(pedestrians$Weather_Conditions)[1])
newData
```

```
##        age    sex Light_Conditions Weather_Conditions
## 1  26 - 35   Male         Daylight Fine no high winds
## 2    0 - 5   Male         Daylight Fine no high winds
## 3   6 - 10   Male         Daylight Fine no high winds
## 4  11 - 15   Male         Daylight Fine no high winds
## 5  16 - 20   Male         Daylight Fine no high winds
## 6  21 - 25   Male         Daylight Fine no high winds
## 7  36 - 45   Male         Daylight Fine no high winds
## 8  46 - 55   Male         Daylight Fine no high winds
## 9  56 - 65   Male         Daylight Fine no high winds
## 10 66 - 75   Male         Daylight Fine no high winds
## 11 Over 75   Male         Daylight Fine no high winds
## 12 26 - 35 Female         Daylight Fine no high winds
## 13   0 - 5 Female         Daylight Fine no high winds
## 14  6 - 10 Female         Daylight Fine no high winds
## 15 11 - 15 Female         Daylight Fine no high winds
## 16 16 - 20 Female         Daylight Fine no high winds
## 17 21 - 25 Female         Daylight Fine no high winds
## 18 36 - 45 Female         Daylight Fine no high winds
## 19 46 - 55 Female         Daylight Fine no high winds
## 20 56 - 65 Female         Daylight Fine no high winds
## 21 66 - 75 Female         Daylight Fine no high winds
## 22 Over 75 Female         Daylight Fine no high winds
```

```
#'
#+ pred
thePred = predict(theGlm, newData, se.fit=TRUE)
thePred2 = do.call(cbind, thePred[1:2])
thePred3 = thePred2 %*% Pmisc::ciMat(0.99)
#'

#+ simplePlot
newData$ageInt= as.numeric(gsub("[[:punct:]].*|[[:alpha:]]", "", newData$age))
```

```r
matplot(newData$ageInt, thePred3, pch = c(16,1,1), col='black')
theMales = newData$sex == 'Male'
matpoints(newData[theMales, 'ageInt'], thePred3[theMales,], pch=c(16,1,1), col='red')
legend('topleft', fill=c('red','black'), legend=c('Male','Female'),bty='n')
```




```r
#'
```

GLMM: killed when armed

```r
# load data
Durl =paste0("https://interactive.guim.co.uk/2015/the-counted/","thecounted-data.zip")
Dfile =file.path(tempdir(),basename(Durl))
if(!file.exists(Dfile))download.file(Durl, Dfile,method ="curl")
unzip(Dfile,exdir =tempdir())
Cfile =file.path(tempdir(),grep("csv$",unzip(Dfile,list =TRUE)$Name,
                                value =TRUE))
deaths =NULL
for(D in Cfile) deaths =rbind(read.table(D,header =TRUE,
                                         sep =",",stringsAsFactors =FALSE), deaths)
```

```r
# reshape the data into long format
dTable =reshape2::dcast(deaths,raceethnicity~armed, fun.aggregate=length)
```

- Does race affect the probability ofbeing killed while unarmed?

```r
Sgender =c("Male","Female")
Srace =c("White","Black","Hispanic/Latino")
Sarmed =c("Firearm","Non-lethal firearm","No")
deathsSub =deaths[deaths$raceethnicity %in% Srace & deaths$armed %in% Sarmed & deaths$gender %in%
deathsSub$raceethnicity =factor(deathsSub$raceethnicity,levels =Srace)
deathsSub$gender =factor(deathsSub$gender,levels =Sgender)
deathsSub$unarmed =as.numeric(deathsSub$armed=="No")

dTable$odds = dTable$No / dTable$Firearm
knitr::kable(dTable[,c('raceethnicity', 'Firearm','No', 'odds' )])
```

raceethnicity	Firearm	No	odds
Arab-American	2	2	1.0000000
Asian/Pacific Islander	14	5	0.3571429
Black	283	121	0.4275618
Hispanic/Latino	161	67	0.4161491
Native American	17	6	0.3529412
Other	1	0	0.0000000
Unknown	18	3	0.1666667
White	564	201	0.3563830

```r
# The glm model

deathResult =glm(unarmed~raceethnicity+gender,
                 family=binomial(link='logit'),data=deathsSub)

knitr::include_graphics("5.png")
```

	est	2.5 %	97.5 %
ref prob			
White:Male	0.22	0.18	0.25
raceethnicity			
Black	1.27	0.97	1.68
Hispanic/Latino	1.19	0.83	1.69
gender			
Female	3.00	1.86	4.83
sd			
state	0.32	0.18	0.56

- Case: killed while armed;

- Control: killed while unarmed.
- Problem with this model: The control group over-represents whites, in order to be killed while armed, one must have a gun, guns are more common in rural areas, and rural areas are more white ….
- Add the State as random effect (gun ownership vary by states, and we assume it doesn't vary within states)!

```
dRes =glmmTMB::glmmTMB(unarmed~raceethnicity+gender+(1|state),
                       data =deathsSub,
                       family =binomial(link ="logit"))
Pmisc::mdTable(Pmisc::coefTable(dRes)$table,guessGroup =F,digits =2)
```

表 2: {#tbl:unnamed-chunk-18}

	variable	level	est	2.5 %	97.5 %
(Intercept)	ref prob	White:Male	0.22	0.18	0.25
raceethnicityBlack	raceethnicity	Black	1.27	0.97	1.68
raceethnicityHispanic/Latino	raceethnicity	Hispanic/Latino	1.19	0.83	1.69
genderFemale	gender	Female	3.00	1.86	4.83
state.SD	sd	state	0.32	0.18	0.56

Conclusion: - biased against black;
- woman are biased due to gun ownership…