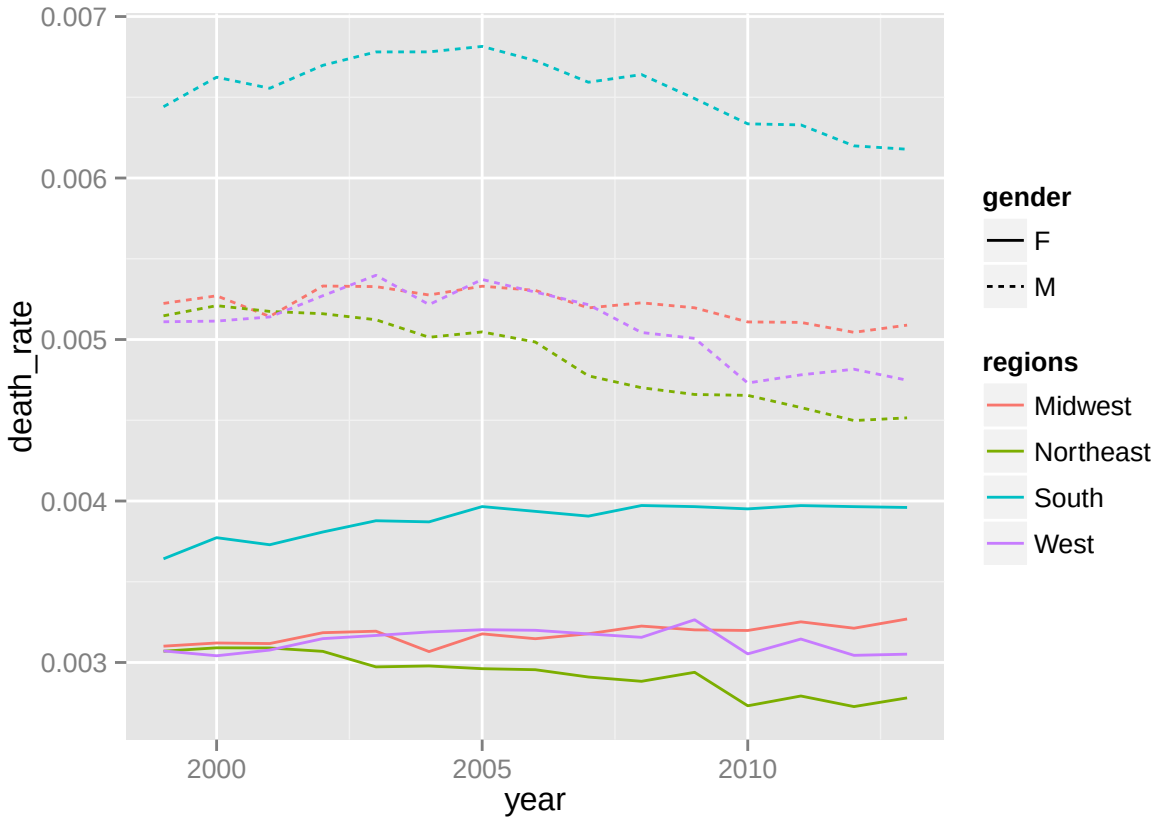
HW10

I started off this project looking at some data that I downloaded from the CDC. In addition to the gender data that everyone had already looked at, I wanted to look at geography. I first plotted region and gender in terms of absolute death rate, showing that the South was significantly worst than every other region while the other three started off roughly the same in 1999 and the Northeast seems to be pulling away in recent years.

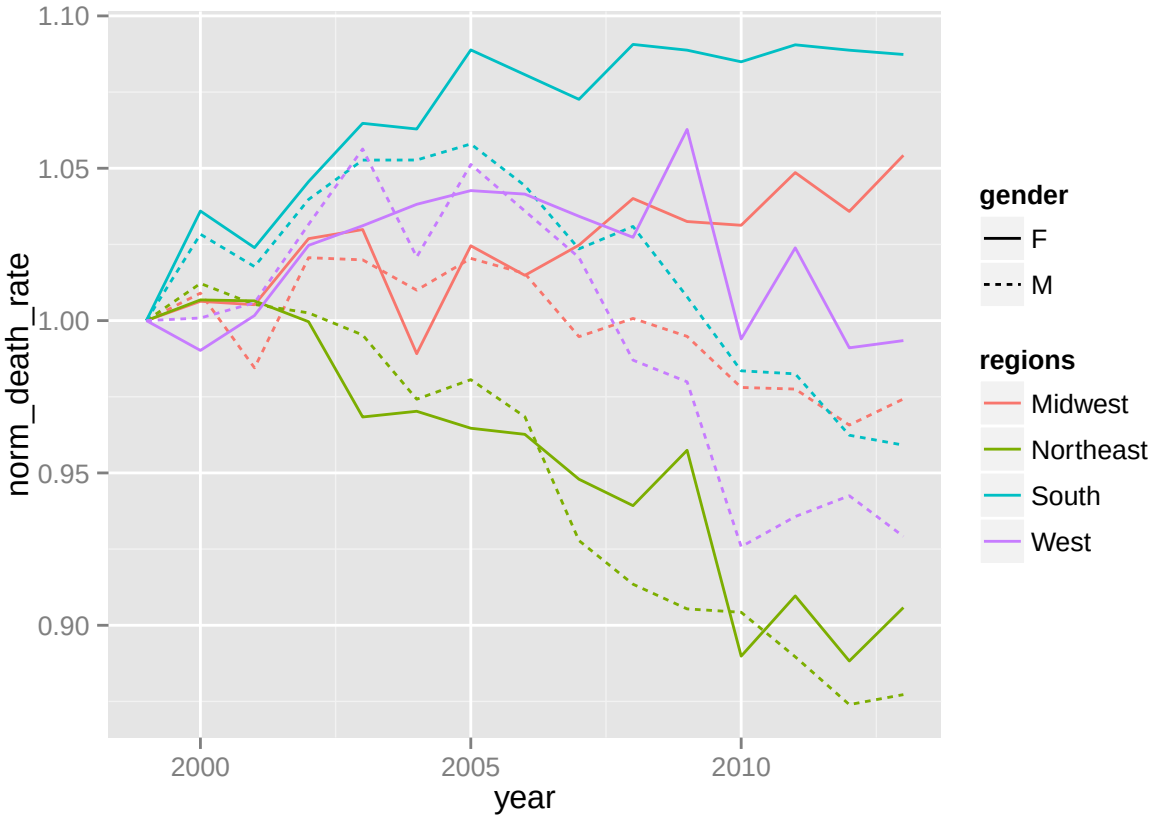
```
library(tidyr)
library(dplyr)
library(ggplot2)
setwd("~/Documents/BDA/Homework 10")
causes_of_death <- read.delim("~/Documents/BDA/Homework 10/causes_of_death.txt")
ok <- causes_of_death["Hispanic.Origin.Code"]=="2186-2"
nhl <- causes_of_death[ok,]

age <- nhl$Single.Year.Ages.Code
year <- as.factor(nhl$Year.Code)
N_years <- nlevels(year)
year <- as.numeric(year)
gender <- as.factor(nhl$Gender.Code)
N_genders <- nlevels(gender)
gender <- as.numeric(gender)
region <- as.factor(nhl$Census.Region.Code)
N_regions <- nlevels(region)
region <- as.numeric(region)
deaths <- nhl$Deaths
population <- nhl$Population
rate <- deaths/population
V <- rate * (1 - rate) / population * 1e10
rate <- rate * 1e5
#rate <- rate - mean(rate)
N <- length(age)

male <- gender=="M"
female <- gender=="F"
R1 <- region=="CENS-R1"
R2 <- region=="CENS-R2"
R3 <- region=="CENS-R3"
R4 <- region=="CENS-R4"
mort_data <- data.frame(age, year, gender, region, deaths, population)

years_1 <- 1999:2013
ages_decade <- list(35:44, 45:54, 55:64)
male_raw_death_rate <- array(NA, length(years_1))
female_raw_death_rate <- array(NA, length(years_1))
avg_death_rate <- array(NA, length(years_1))
male_avg_death_rate <- array(NA, length(years_1))
female_avg_death_rate <- array(NA, length(years_1))
R1_avg_death_rate <- array(NA, length(years_1))
R2_avg_death_rate <- array(NA, length(years_1))
R3_avg_death_rate <- array(NA, length(years_1))
R4_avg_death_rate <- array(NA, length(years_1))
```

```
R1_male_avg_death_rate <- array(NA, length(years_1))
R2_male_avg_death_rate <- array(NA, length(years_1))
R3_male_avg_death_rate <- array(NA, length(years_1))
R4_male_avg_death_rate <- array(NA, length(years_1))
R1_female_avg_death_rate <- array(NA, length(years_1))
R2_female_avg_death_rate <- array(NA, length(years_1))
R3_female_avg_death_rate <- array(NA, length(years_1))
R4_female_avg_death_rate <- array(NA, length(years_1))

data <- nhl
male <- data[,"Gender.Code"]=="M"
R1 <- data[,"Census.Region.Code"]=="CENS-R1"
R2 <- data[,"Census.Region.Code"]=="CENS-R2"
R3 <- data[,"Census.Region.Code"]=="CENS-R3"
R4 <- data[,"Census.Region.Code"]=="CENS-R4"
for (i in 1:length(years_1)){
  ok <- data[,"Year"]==years_1[i] & data[,"Single.Year.Ages.Code"] %in% ages_decade[[2]]
  avg_death_rate[i] <- mean(data[ok,"Deaths"]/data[ok,"Population"])
  male_avg_death_rate[i] <- mean(data[ok&male,"Deaths"]/data[ok&male,"Population"])
  female_avg_death_rate[i] <- mean(data[ok&!male,"Deaths"]/data[ok&!male,"Population"])
  R1_avg_death_rate[i] <- mean(data[ok&R1,"Deaths"]/data[ok&R1,"Population"])
  R2_avg_death_rate[i] <- mean(data[ok&R2,"Deaths"]/data[ok&R2,"Population"])
  R3_avg_death_rate[i] <- mean(data[ok&R3,"Deaths"]/data[ok&R3,"Population"])
  R4_avg_death_rate[i] <- mean(data[ok&R4,"Deaths"]/data[ok&R4,"Population"])
  R1_male_avg_death_rate[i] <- mean(data[ok&R1&male,"Deaths"]/data[ok&R1&male,"Population"])
  R2_male_avg_death_rate[i] <- mean(data[ok&R2&male,"Deaths"]/data[ok&R2&male,"Population"])
  R3_male_avg_death_rate[i] <- mean(data[ok&R3&male,"Deaths"]/data[ok&R3&male,"Population"])
  R4_male_avg_death_rate[i] <- mean(data[ok&R4&male,"Deaths"]/data[ok&R4&male,"Population"])
  R1_female_avg_death_rate[i] <- mean(data[ok&R1&!male,"Deaths"]/data[ok&R1&!male,"Population"])
  R2_female_avg_death_rate[i] <- mean(data[ok&R2&!male,"Deaths"]/data[ok&R2&!male,"Population"])
  R3_female_avg_death_rate[i] <- mean(data[ok&R3&!male,"Deaths"]/data[ok&R3&!male,"Population"])
  R4_female_avg_death_rate[i] <- mean(data[ok&R4&!male,"Deaths"]/data[ok&R4&!male,"Population"])
}
year <- rep(1999:2013, 8)
regions <- rep(c(rep("Northeast", 15), rep("Midwest", 15), rep("South", 15), rep("West", 15)), 2)
gender <- c(rep("M", 60), rep("F", 60))
death_rate <- c(R1_male_avg_death_rate, R2_male_avg_death_rate, R3_male_avg_death_rate, R4_male_avg_dea
norm_death_rate <- c(R1_male_avg_death_rate/R1_male_avg_death_rate[1], R2_male_avg_death_rate/R2_male_a

df <- data.frame(year, regions, gender, death_rate)
norm_df <- data.frame(year, regions, gender, norm_death_rate)
ggplot(data = df, aes(x = year, y = death_rate, group = interaction(gender, regions), color = regions,
```



This effect is even starker if we look at relative death rates:

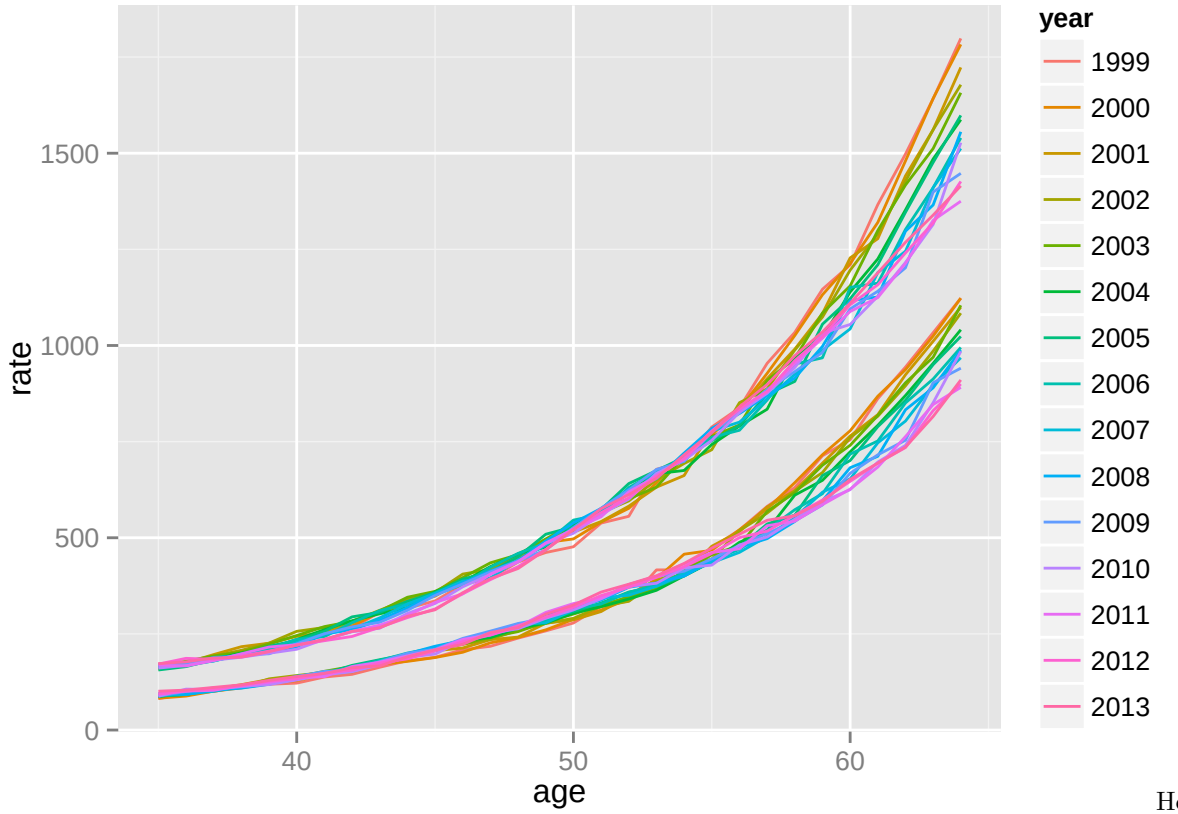
```
ggplot(data = norm_df, aes(x = year, y = norm_death_rate, group = interaction(gender, regions), color =
```



It seems that death rates among non-hispanic whites have been consistently dropping in the Northeast (perhaps on par with the rest of the world) and holding steady in the Midwest and West. However, the South seems to be driving a lot of the effects that Andrew posted about with the death rates for mean dropping sharply after 2005 and death rates for woman continuing to shoot up (although perhaps recently plateauing). In order to investigate this further, I tried to make a regression model. Because I had now significantly stratified the data, there were now very few deaths in some of the cells in my dataset. Therefore, I had a lot of uncertainty about the exact estimates of the death rates in each gender/year/region interaction. I thus tried to model death rates directly and include my uncertainty by adding a term for the error in the rate estimate which was $V(i) = p[i](1 - p[i])/N[i]$. I originally tried to use a binomial logit model but then switched to a normal model to make the coefficients easier to interpret. In order to do this, I used the rate per 10,000 (and so had to multiply the error accordingly). I started building my model but quickly ran into problems with the linear model. This was because the relationship between age and rate was clearly closer to an exponential.

```
library(rstan)
death_rates <- read.delim("~/Documents/BDA/Homework 10/cdc/white_nonhisp_death_rates_from_1999_to_2013_

age <- death_rates$Age
year <- as.factor(death_rates$Year)
N_years <- nlevels(year)
year <- as.numeric(year)
gender <- as.factor(death_rates$Male)
N_genders <- nlevels(gender)
gender <- as.numeric(gender)
deaths <- death_rates$Deaths
population <- death_rates$Population
rate <- deaths/population
V <- rate * (1 - rate) / population * 1e10
rate <- rate * 1e5
```

```r
N <- length(age)
decades <- rep(NA, length(age))
decades[age %in% 35:44] <- 1
decades[age %in% 45:54] <- 2
decades[age %in% 55:65] <- 3
N_decades <- 3

year <- as.factor(death_rates$Year)
ggplot(data = death_rates, aes(x = age, y = rate, group = interaction(gender, year), color = year)) + g
```



However, I wanted to keep things linear because I wanted to maintain my estimate of the error in the rate calculation. My final model was rate $\sim \beta_a$ (age) + β_y (year) + β_g (gender) + $\beta_y d$ (year * generation) + constant (where generation stood for the age ranges used in the Case and Deaton study). I wanted to include hierarchical priors but I was only able to include one for year. This seemed to be because the effects for gender were so large (around +80, -80) that it was taking the sampler too long to find reasonable values for the hierarchical variance parameter. Stan code for this model:

```
data {
  int<lower=0> N;
  int<lower=0> N_years;
  int<lower=0> N_genders;
  real V[N];
  real rate[N];
  vector[N] age;
  int year[N];
  int gender[N];
  int decades[N];
}
```

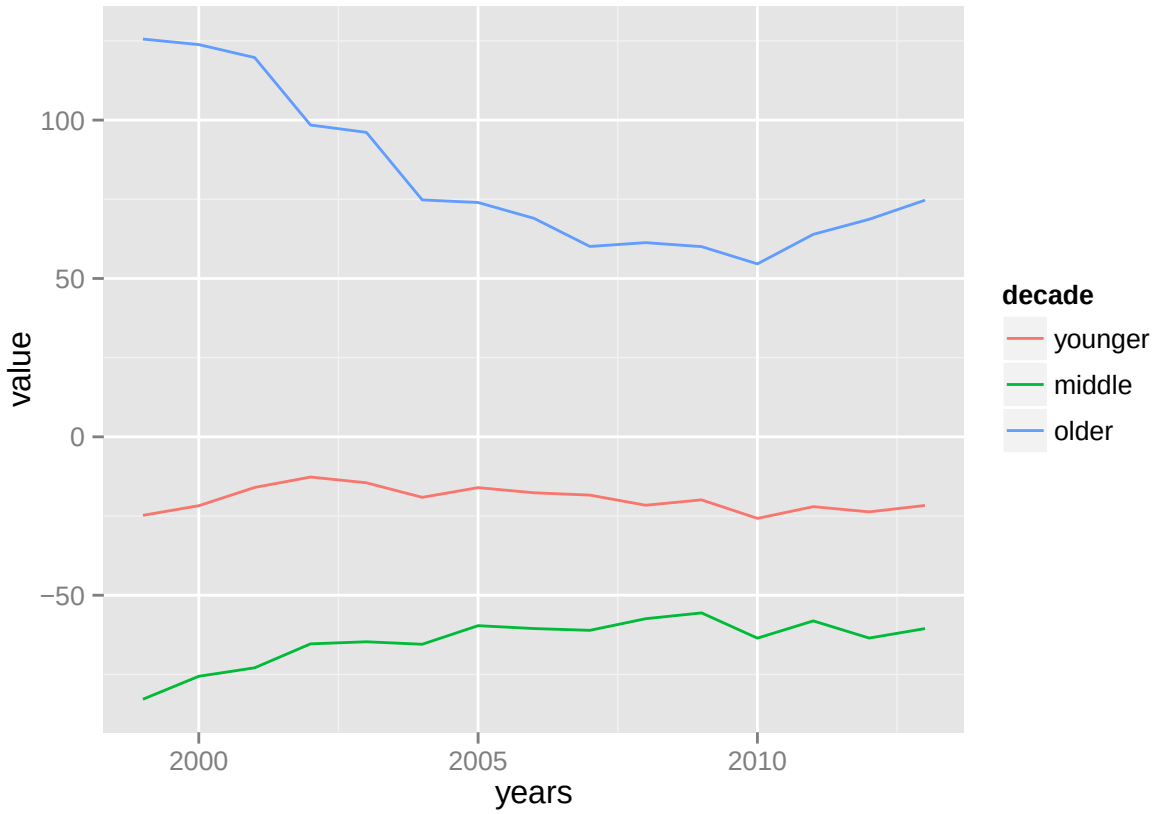
```
parameters {
  real<lower=0> sigma_sq;
  real constant;
  real beta_a;
  vector[N_years] beta_y;
  vector[N_genders] beta_g;
  vector[N_years] beta_y1;
  vector[N_years] beta_y2;
  vector[N_years] beta_y3;
  real<lower=0> tau_y;
}
transformed parameters {
  vector[N_years] beta_yd[3];
  beta_yd[1] <- beta_y1;
  beta_yd[2] <- beta_y2;
  beta_yd[3] <- beta_y3;
}
model {
  beta_y ~ normal(0, tau_y);
  beta_g ~ normal(0, 100);
  beta_a ~ normal(30, 30);
  beta_y1 ~ normal(0, 100);
  beta_y2 ~ normal(0, 100);
  beta_y3 ~ normal(0, 100);
  for(i in 1:N){
    rate[i] ~ normal(beta_a * age[i] + beta_y[year[i]] + beta_g[gender[i]] + beta_yd[decades[i], year[i]
  }
}
```

I now fit the model, subtracting the mean of the rate and age first (mostly for the upcoming Gaussian process model).

```
rate <- rate - mean(rate)
age <- age - 35
year <- as.numeric(year)
fit <- stan("simple_model.stan", iter = 1000)
m <- as.data.frame(monitor(fit))
```

However, when I looked at the interaction terms (which should have shown similar effects to the plots from Andrew's blog post), they showed something entirely different:

```
means <- select(m, `mean`)
years <- 1999:2013
younger <- means[21:35,]
middle <- means[36:50,]
older <- means[51:65,]
df <- data.frame(years, younger, middle, older)
df <- gather(df, decade, value, younger, middle, older)
ggplot(data = df, aes(x = years, y = value, group = decade, color = decade)) + geom_line()
```



It seems that almost all of the coefficients effect was to correct for the linear assumption about the relationship between age and rate, which is why the 45-54 year olds show highly negative values. Therefore, I decided to fit a Gaussian process model for age and year in order to properly characterize that relationship and then build that into my full regression. For this, I used the model $\begin{pmatrix} y \\ \tilde{\mu} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x)+\sigma^2 I & K(\tilde{x},x)) \\ K(x,\tilde{x})) & K(\tilde{x},\tilde{x})) \end{pmatrix}$. where y is the age/year effect on the death rate. This Gaussian process is actually a sum of two Gaussian processes, one for age and one for year. Thus we can write for any data point d (an age), $f_{\text{male}}(d) = GP(0, k_{\text{age}})$ where $k_{\text{age}}(d, d') = \tau^2_{\text{male}} \exp(\frac{(d_{\text{age}} - d'_{\text{age}})^2}{l^2_{\text{male}}})$ with the one for year being exactly similar. The stan code follows:

```
data {
  int<lower=0> N;
  int<lower=0> N_years;
  int<lower=0> N_genders;
  real V[N];
  vector[N] rate;
  vector[N] age;
  int year[N];
  int gender[N];
  int decades[N];
}
transformed data {
  vector[N] mu;
  for (i in 1:N)
    mu[i] <- 0;
}
parameters {
  real<lower=0> sigma_sq;
  real constant;
```

```
  real<lower=0> tau_sq_a;
  real<lower=0> l_sq_a;
  real<lower=0> tau_sq_y;
  real<lower=0> l_sq_y;
  real<lower=0> sigma_sq_ay;
  vector[N] beta_ay;
}
model {
  matrix[N,N] Sigma;
  matrix[N,N] Sigma_a;
  matrix[N,N] Sigma_y;
  // off-diagonal elements
  for (i in 1:N) {
    for (j in 1:N) {
      Sigma_a[i,j] <- (tau_sq_a * exp(-pow(age[i] - age[j],2)/l_sq_a) + if_else(i==j, sigma_sq + V[i], 
      Sigma_y[i,j] <- (tau_sq_y * exp(-pow(year[i] - year[j],2)/l_sq_y) + if_else(i==j, 0.0, 0.0));
    }
  }
  Sigma <- Sigma_a + Sigma_y;

  tau_sq_a ~ normal(0,5);
  l_sq_a ~ normal(0,5);
  tau_sq_y ~ normal(0,5);
  l_sq_y ~ normal(0,5);

  //beta_ay ~ multi_normal(mu,Sigma);
  //for(i in 1:N){
    //rate[i] ~ normal(beta_a * age[i] + beta_y[year[i]] + beta_g[gender[i]] + beta_yd[decades[i], year
  //}
  rate ~ multi_normal(mu,Sigma);
}
```

I wanted to continue building this full model but the sampler did not seem to work for this Gaussian process.
I spent a couple days trying to figure this out but could not and therefore do not have any further modeling.