HW7

a. We can write the model used for this Gaussian process as $\begin{pmatrix} y \\ \tilde{\mu} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x)+\sigma^2 I & K(\tilde{x},x)) \\ K(x,\tilde{x})) & K(\tilde{x},\tilde{x})) \end{pmatrix})$. This Gaussian process is actually a sum of two Gaussian processes, one for each gender. Thus we can write for any data point d (an age and gender), $f_{\text{male}}(d) = GP(0, k_{\text{male}})$ where $k_{\text{male}}(d,d') = I_{\text{male}}(d_{\text{gender}})\tau^2_{\text{male}} \exp(\frac{(d_{\text{age}}-d'_{\text{age}})^2}{l^2_{\text{male}}})$. The one for female is exactly the same. We use a linear probability model here for ease of interpretation since none of the data are close to 0 or 1. When a data point d is in the dataset, we can use the actual variance of the proportion estimate which is $p(1-p)/N$

b. A function for computing the log posterior:

```
log_posterior <- function(age_data, gender_data, tau_sq_f, tau_sq_m, l_sq_f, l_sq_m, sigma_sq) {
  N_data <- length(ages)
  Sigma_m <- matrix(, nrow = N_data, ncol = N_data)
  Sigma_f <- matrix(, nrow = N_data, ncol = N_data)
  Sigma <- matrix(, nrow = N_data, ncol = N_data)
  for (i in 1:N_data-1) {
    for (j in i+1:N_data) {
      Sigma_m[i,j] <- (1 - gender_data[i]) * (1 - gender_data[j]) *
        (tau_sq_m * exp((-(ageBin_data[i] - ageBin_data[j])^2)/l_sq_m))
      Sigma_f[i,j] <- gender_data[i] * gender_data[j] *
        (tau_sq_f * exp((-(ageBin_data[i] - ageBin_data[j])^2)/l_sq_f))
    }
  }
  Sigma <- diag(sigma_sq)
  Sigma <- Sigma + Sigma_m + Sigma_f
  dmvnorm(age_data, mean = rep(0, N_data), sigma = Sigma, log = TRUE)
}
```

c. The model in Stan:

```
data {
  int<lower=1> N_data;
  vector[N_data] ageBin_data;
  vector[N_data] gks_data;
  vector[N_data] gender_data;
  vector[N_data] sigma_sq;
}
transformed data {
  vector[N_data] mu;
  for (i in 1:N_data)
    mu[i] <- 0;
}
parameters {
  real<lower=0> tau_sq_m;
  real<lower=0> l_sq_m;
  real<lower=0> tau_sq_f;
  real<lower=0> l_sq_f;
}
model {
  matrix[N_data,N_data] Sigma_m;
  matrix[N_data,N_data] Sigma_f;
  matrix[N_data,N_data] Sigma;

  tau_sq_m ~ cauchy(0,5);
  l_sq_m ~ cauchy(0,5);
  tau_sq_f ~ cauchy(0,5);
  l_sq_f ~ cauchy(0,5);

  // off-diagonal elements
  for (i in 1:N_data) {
    for (j in 1:N_data) {
      Sigma_m[i,j] <- (1 - gender_data[i]) * (1 - gender_data[j]) *
        (tau_sq_m * exp(-pow(ageBin_data[i] - ageBin_data[j],2)/l_sq_m) + if_else(i==j, sigma_sq[i], 0.0
      Sigma_f[i,j] <- gender_data[i] * gender_data[j] *
        (tau_sq_f * exp(-pow(ageBin_data[i] - ageBin_data[j],2)/l_sq_f) + if_else(i==j, sigma_sq[i], 0.0
    }
  }

  Sigma <- Sigma_m + Sigma_f;

  gks_data ~ multi_normal(mu,Sigma);
}
```

We now fit the model

```r
library(rstan)
library(dplyr)
library(ggplot2)
library(tidyr)
library(mvtnorm)
setwd("~/Documents/BDA/Homework 7")
naes04 <- read.csv("naes04.csv")
naes04$ageBin <- cut(naes04$age, 16, labels = c(seq(20, 95, 5)))
by_age_gender <- group_by(naes04, ageBin, gender)
pp_naes <- summarise(by_age_gender,
          count = n(),
          gks = mean(as.numeric(gayKnowSomeone) - 1, na.rm = TRUE))
pp_naes$sigma_sq <- (pp_naes$gks * (1 - pp_naes$gks))/pp_naes$count
ageBin_data <- as.numeric(as.character(pp_naes$ageBin))[1:32]
gks <- pp_naes$gks[1:32]
gks_data <- gks - mean(gks)
sigma_sq <- pp_naes$sigma_sq[1:32]
gender_data <- as.numeric(pp_naes$gender)[1:32] - 1
N_data <- length(gks_data)
fit <- stan(file="gp.stan")
```

d. And now to visualize the uncertainty, we take several draws from the model using the mean fit parameters:

```r
samps <- rstan::extract(fit, permuted = TRUE)
tau_sq_m <- mean(samps$tau_sq_m)
tau_sq_f <- mean(samps$tau_sq_f)
l_sq_m <- mean(samps$l_sq_m)
```

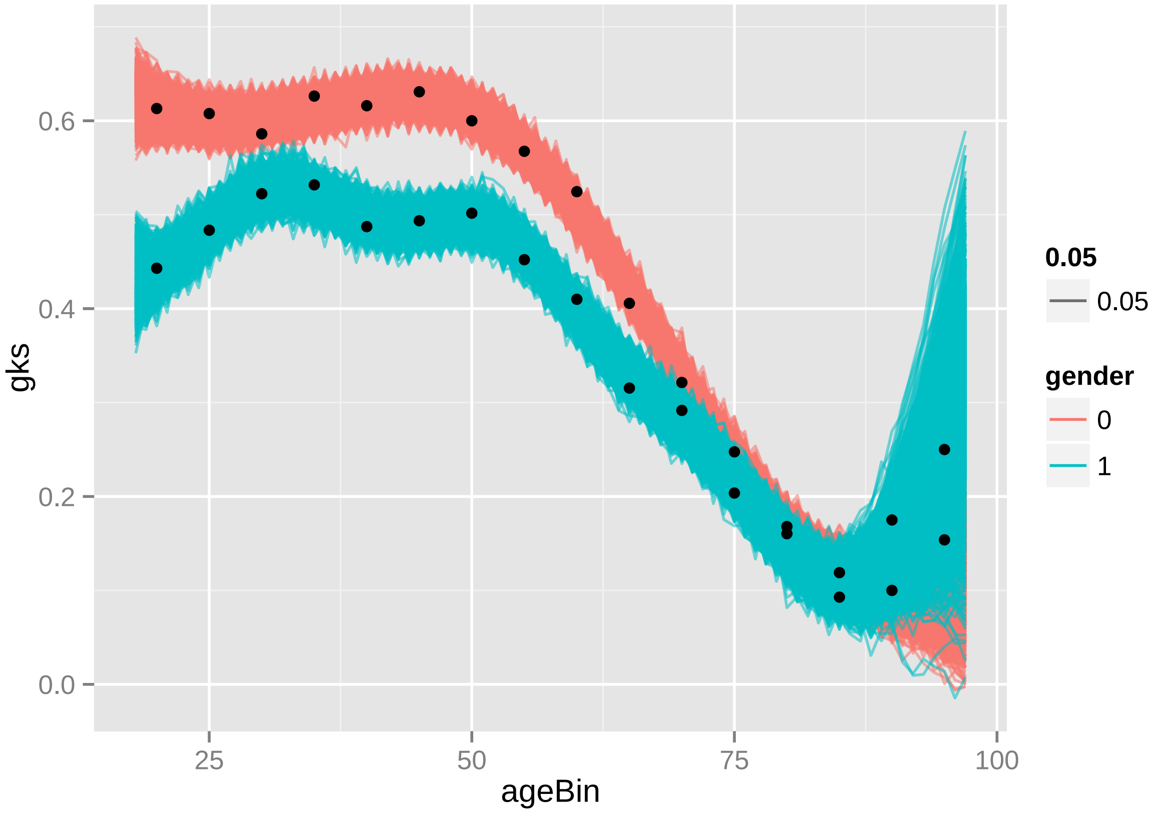
```
l_sq_f <- mean(samps$l_sq_f)

ageBin_sim <- c(18:97, 18:97);
N_sim <- length(ageBin_sim);
gender_sim <- c(rep(0, N_sim/2), rep(1, N_sim/2));

fit_predict <- stan(file="gp-predict.stan");
fit_predict_ss <- rstan::extract(fit_predict, permuted=TRUE);
df <- data.frame(x_m=ageBin_sim, y_sim_m=colMeans(fit_predict_ss$gks)+mean(gks));
sims <- as.data.frame(t(fit_predict_ss$gks+mean(gks)));
sims$ageBin <- ageBin_sim;
sims$gender <- as.factor(gender_sim);
sims_df <- gather(sims, sim, gks, -ageBin, -gender)
data_df <- data.frame(gks=gks_data+mean(gks), ageBin=ageBin_data, gender=gender_data)
```

And plot them:

```
p <- ggplot()
p <- p + geom_line(data=sims_df, aes(x=ageBin, y=gks, color=gender, group=interaction(sim, gender), alph
p <- p + geom_point(data=data_df, aes(x=ageBin, y=gks))
p
```


gks
ageBin
gender