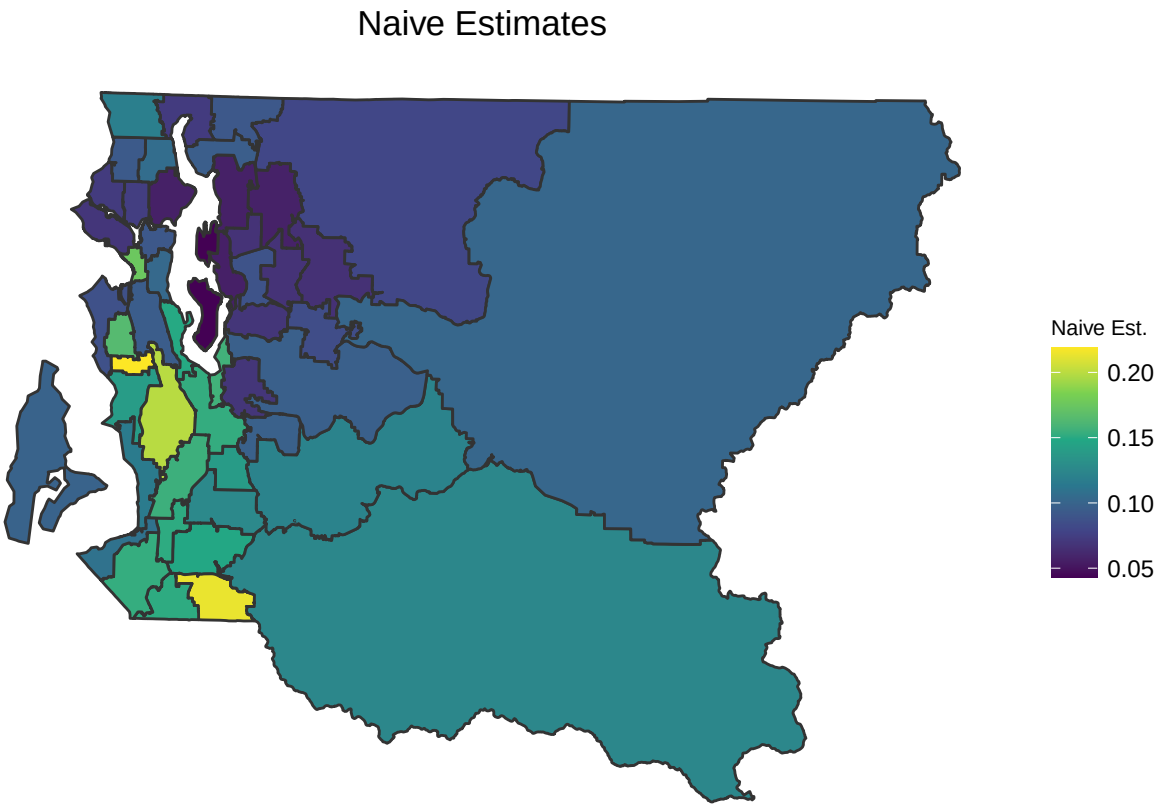
554-2022-Assignment3 key

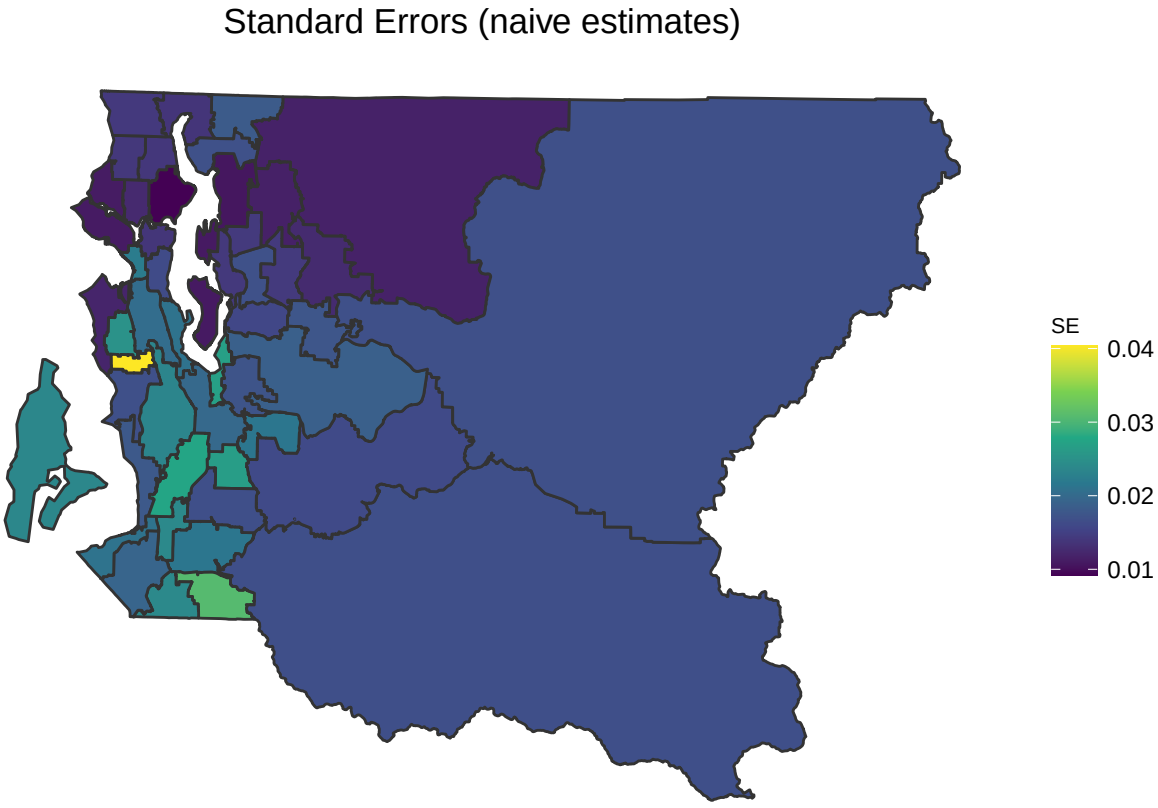
Question 1

```
### load libraries
library(RColorBrewer)
library(rgdal)
library(INLA)
library(SUMMER)
library(spdep)
library(dplyr)
library(survey)
library(ggplot2)

# load data
data(BRFSS)
data(KingCounty)

# create neighborhood matrix for later
nb.r <- poly2nb(KingCounty, queen = F, row.names = KingCounty$HRA2010v2_)
mat <- nb2mat(nb.r, style = "B", zero.policy = TRUE)
colnames(mat) <- rownames(mat)
mat <- as.matrix(mat[1:dim(mat)[1], 1:dim(mat)[1]])

# remove records with missing hracode and smoker1
BRFSS <- subset(BRFSS, !is.na(BRFSS$smoker1))
BRFSS <- subset(BRFSS, !is.na(BRFSS$hracode))

# compute naive estimates and standard errors
naive_data <- BRFSS %>%
  group_by(hracode) %>%
  summarise(est = mean(smoker1),
            n = length(smoker1)) %>%
  mutate(se = sqrt( est * (1 - est)/n)) %>%
  as.data.frame()
# mapPlot() doesn't handle tibbles well, so if you're using tidyverse functionality,
# convert your tibbles back to dataframes before plotting

# map of estimates
mapPlot(data = naive_data, geo = KingCounty, variables = "est",
    by.data = "hracode", by.geo = "HRA2010v2_", legend.label = "Naive Est.",
    is.long = FALSE) +
   ggtitle("Naive Estimates") +
   theme(strip.background = element_blank(),
         plot.title = element_text(hjust = 0.5),
```

```
        panel.border = element_blank(),
        strip.text.x = element_blank())
```




```
# everything inside of theme() are all just extra ggplot options that
# aren't strictly necessary but can be used to make plots look a bit different

# map of standard errors
mapPlot(data = naive_data, geo = KingCounty, variables = "se",
    by.data = "hracode", by.geo = "HRA2010v2_", legend.label = "SE",
    is.long = FALSE) +
   ggtitle("Standard Errors (naive estimates)") +
   theme(strip.background = element_blank(),
        plot.title = element_text(hjust = 0.5),
        panel.border = element_blank(),
        strip.text.x = element_blank())
```



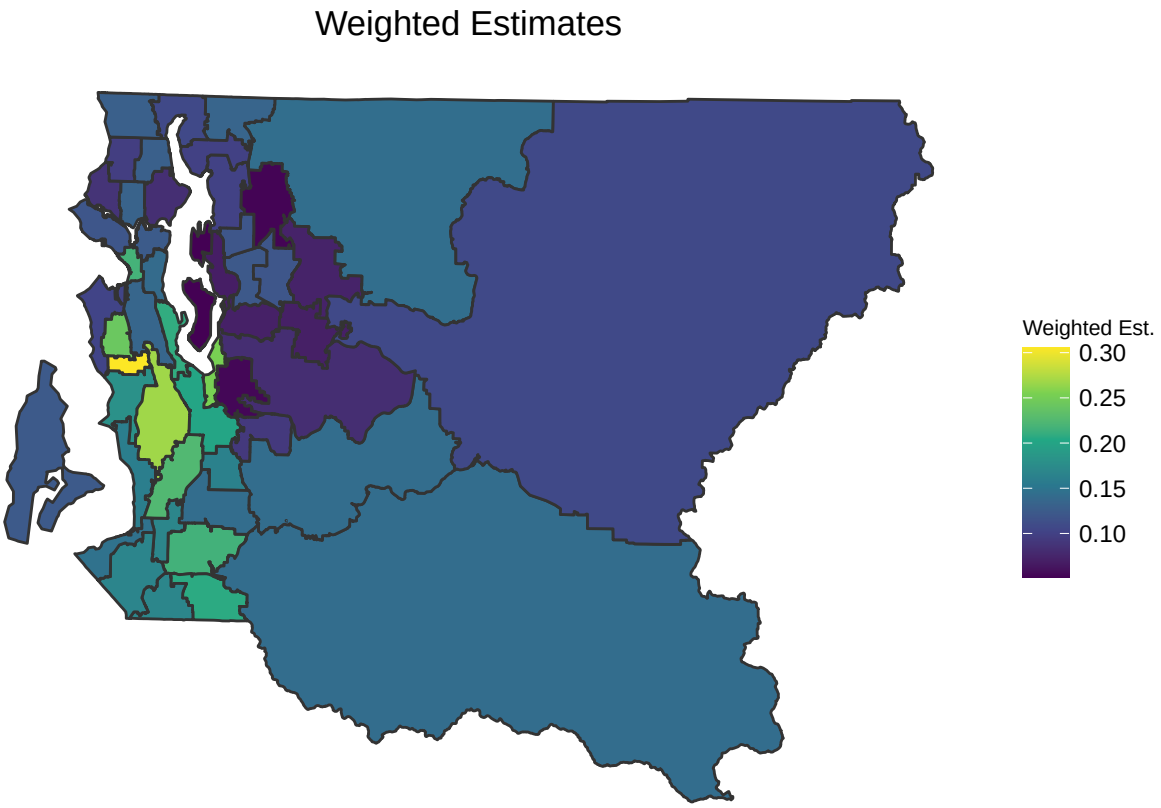
Question 2

```
# create design object, get direct estimates
design <- svydesign(ids = ~1, weights = ~rwt_llcp, strata = ~strata, data = BRFSS)
direct <- svyby(~smoker1, ~hracode, design, svymean)
head(direct, n = 7)
```

```
##                                                hracode    smoker1         se
## Auburn-North                             Auburn-North 0.21603170 0.04928255
## Auburn-South                             Auburn-South 0.20680968 0.04534051
## Ballard                                       Ballard 0.08735027 0.02375056
## Beacon/Gtown/S.Park                Beacon/Gtown/S.Park 0.13550765 0.03539402
## Bear Creek/Carnation/Duvall Bear Creek/Carnation/Duvall 0.14357641 0.03095033
## Bellevue-Central                     Bellevue-Central 0.12307351 0.03189524
## Bellevue-NE                               Bellevue-NE 0.11979605 0.03117550
```

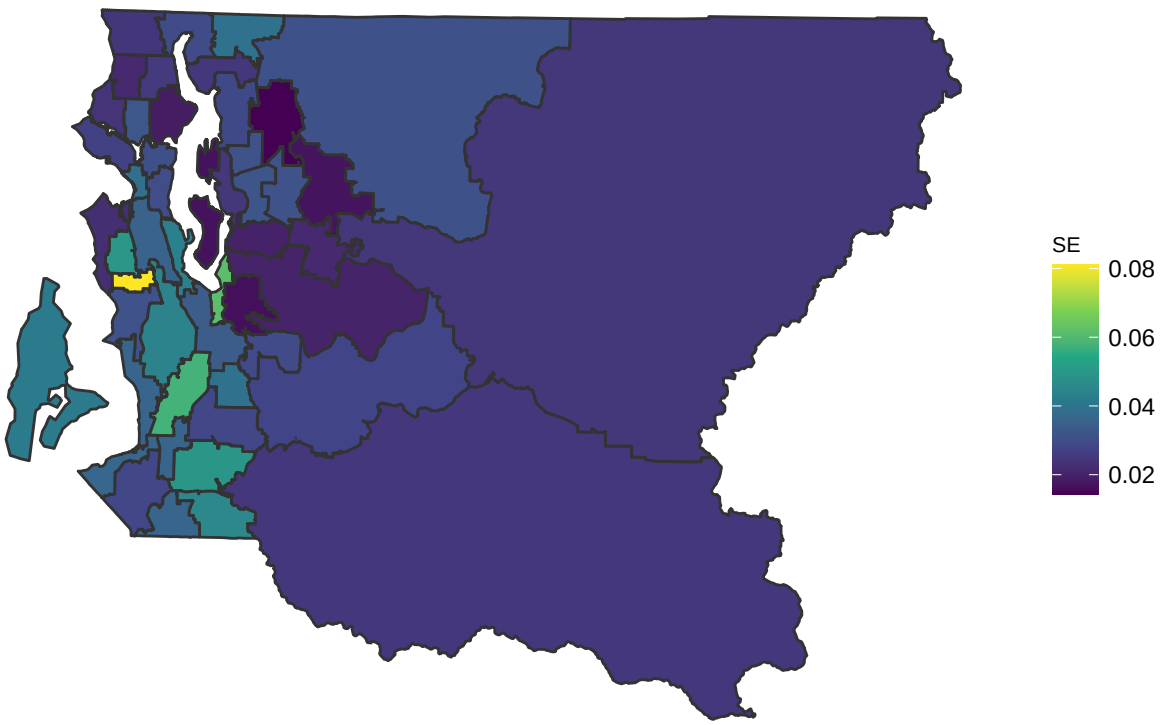
```
# map of weighted estimates
mapPlot(data = direct, geo = KingCounty, variables = "smoker1",
    by.data = "hracode", by.geo = "HRA2010v2_", legend.label = "Weighted Est.",
    is.long = FALSE) +
   ggtitle("Weighted Estimates") +
   theme(strip.background = element_blank(),
         plot.title = element_text(hjust = 0.5),
         panel.border = element_blank(),
         strip.text.x = element_blank())
```

```
## Using hracode as id variables
```



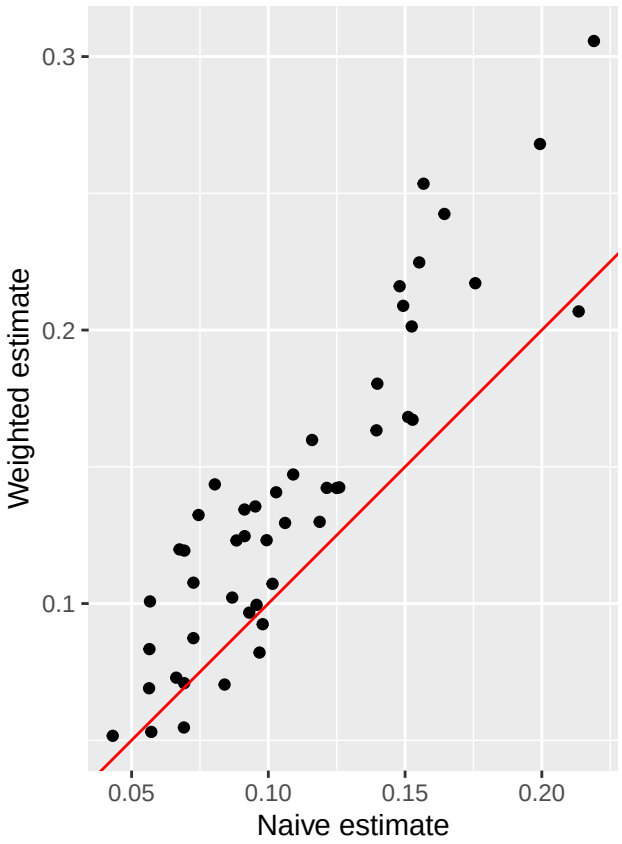

```
# map of standard errors
mapPlot(data = direct, geo = KingCounty, variables = "se",
    by.data = "hracode", by.geo = "HRA2010v2_", legend.label = "SE",
    is.long = FALSE) +
   ggtitle("Standard Errors (weighted estimates)") +
   theme(strip.background = element_blank(),
         plot.title = element_text(hjust = 0.5),
         panel.border = element_blank(),
         strip.text.x = element_blank())
```

```
## Using hracode as id variables
```



Question 3

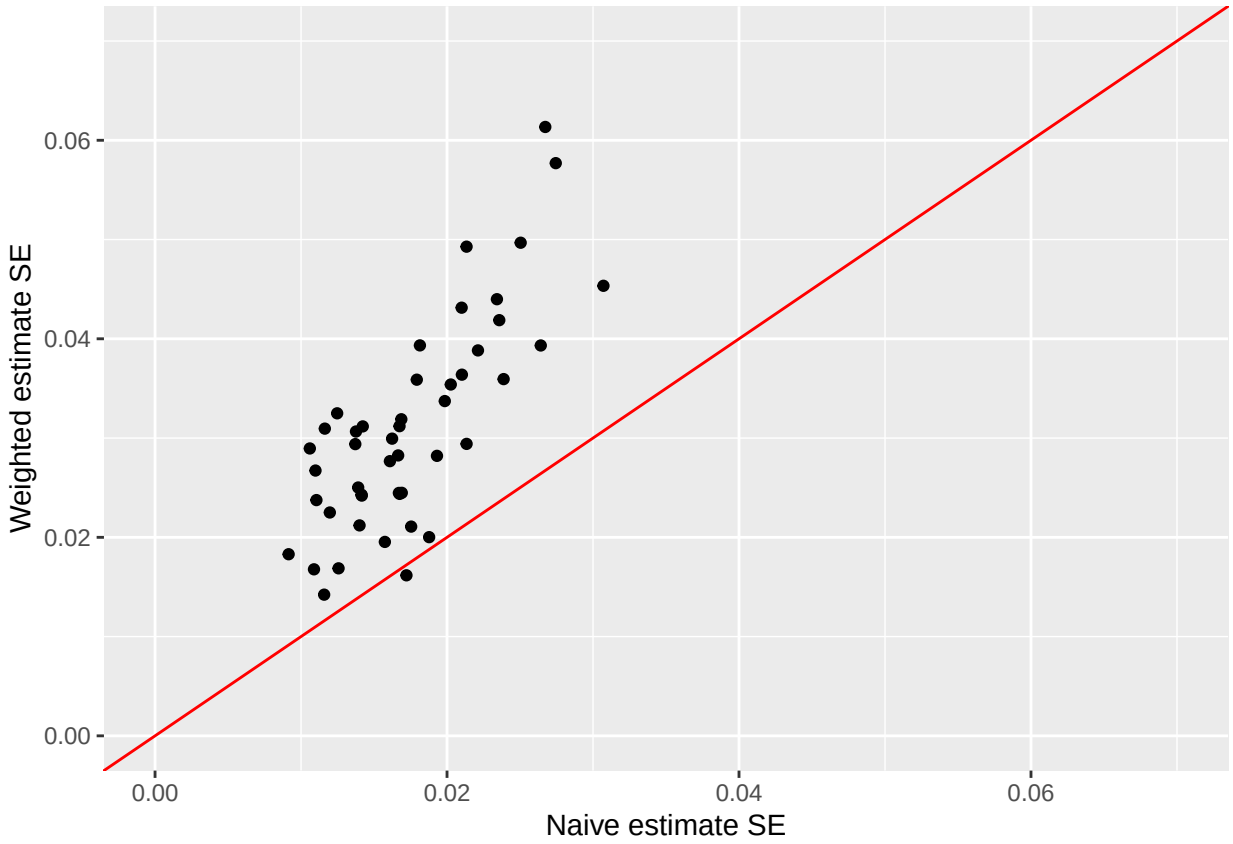
```
# plot point estimates against each other
data.frame(direct = direct$smoker1, naive = naive_data$est) %>%
  ggplot(aes(naive, direct)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red") +
  coord_fixed() +
  xlab("Naive estimate") +
  ylab("Weighted estimate")
```



The weighted estimates are in general higher for most regions than the naive estimates, as evidenced by the majority of the points in the graph above falling above the line with slope = 1, intercept = 0.

```
# plot standard errors against each other
data.frame(direct = direct$se, naive = naive_data$se) %>%
  ggplot(aes(naive, direct)) +
  geom_point() +
  xlab("Naive estimate SE") +
  ylab("Weighted estimate SE") +
  xlim (0, 0.07) +
  ylim (0, 0.07) +
  geom_abline(slope = 1, intercept = 0, col = "red")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The standard errors of the weighted estimates are larger than the standard errors of the naive estimates, possibly because of the higher weighted estimates $\hat{p}_i's$. The standard errors of the naive estimates range from around 0.009 to 0.040, whereas the standard errors of the weighted estimates range from 0.014 to 0.08.

We have stratified, disproportionate sampling in the BRFSS data, and as such estimates that account for the survey design are most appropriate. This means the weighted estimates (and their standard errors) are the more appropriate summaries of the data.

Question 4

```
# smoothed naive estimation
smoothed <- smoothSurvey(data = BRFSS, geo = KingCounty, Amat = mat,
    responseType = "binary", responseVar = "smoker1", strataVar = NULL,
    weightVar = NULL, regionVar = "hracode", clusterVar = NULL,
    CI = 0.95)
```

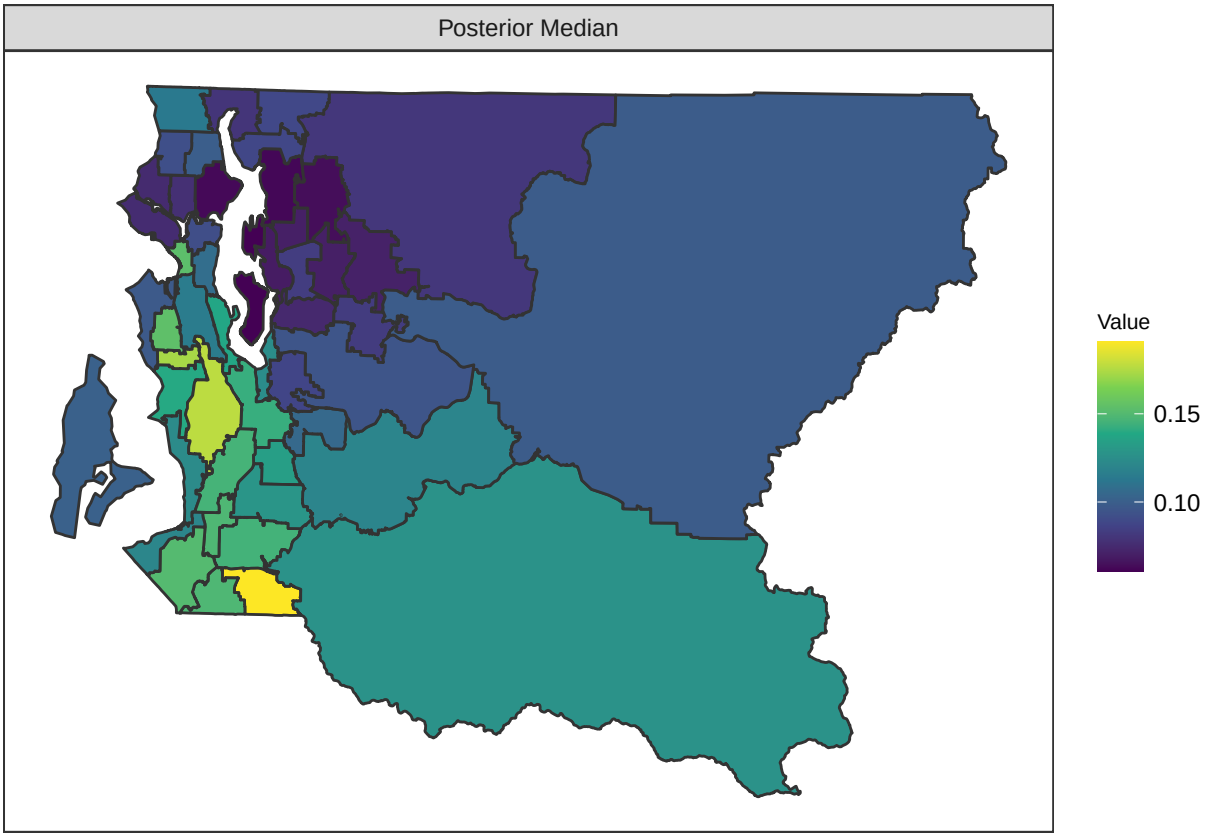
```
## Strata not defined. Ignoring sample design
```

```
## cluster not specified. Ignoring sample design
```

```
## Warning in inla.model.properties.generic(inla.trim.family(model), mm[names(mm) == : Model 'bym2' in s
##   Use this model with extra care!!! Further warnings are disabled.
```

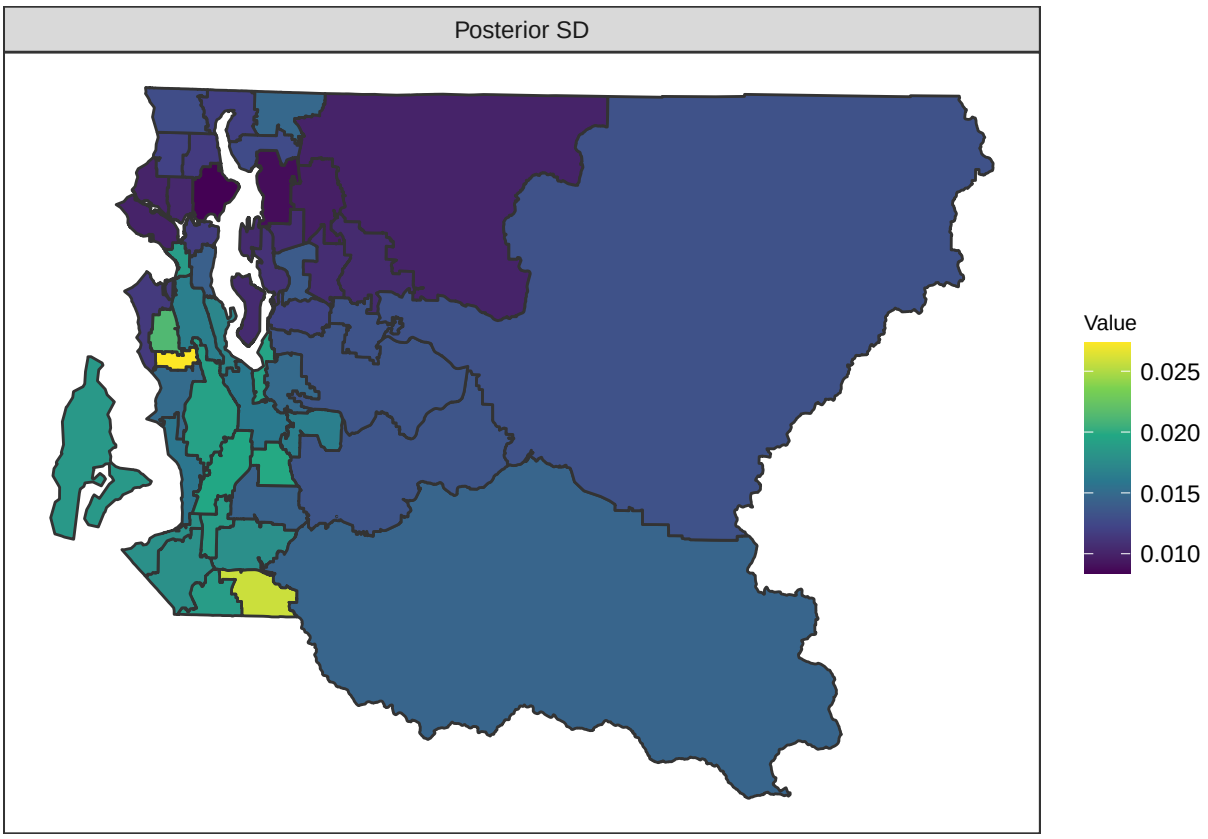
```
# extract posterior medians and map them
toplot <- smoothed$smooth
mapPlot(data = toplot, geo = KingCounty, variables = c("median"),
labels = c("Posterior Median"), by.data = "region", by.geo = "HRA2010v2_")
```

```
## Using region as id variables
```



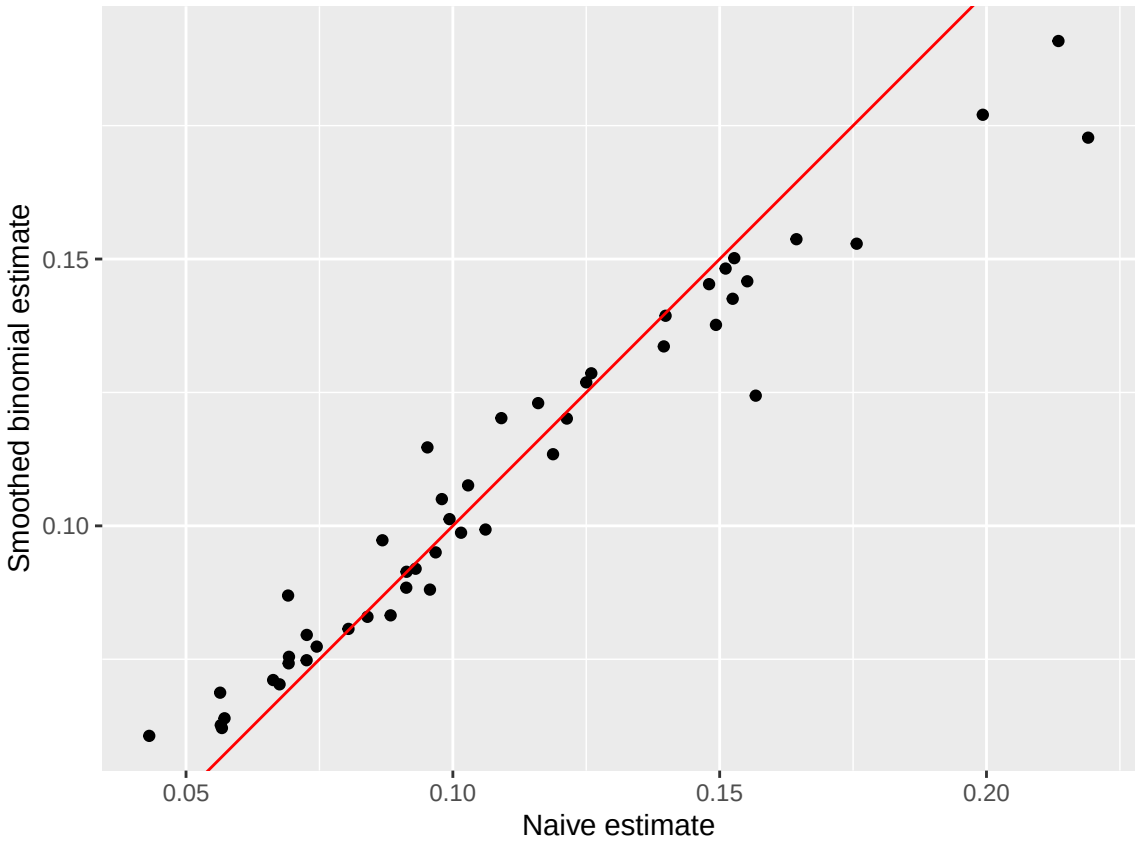

```
# extract posterior standard deviations and map them
toplot$sd <- sqrt(toplot$var)
mapPlot(data = toplot, geo = KingCounty, variables = c("sd"),
labels = c("Posterior SD"), by.data = "region", by.geo = "HRA2010v2_")
```

```
## Using region as id variables
```



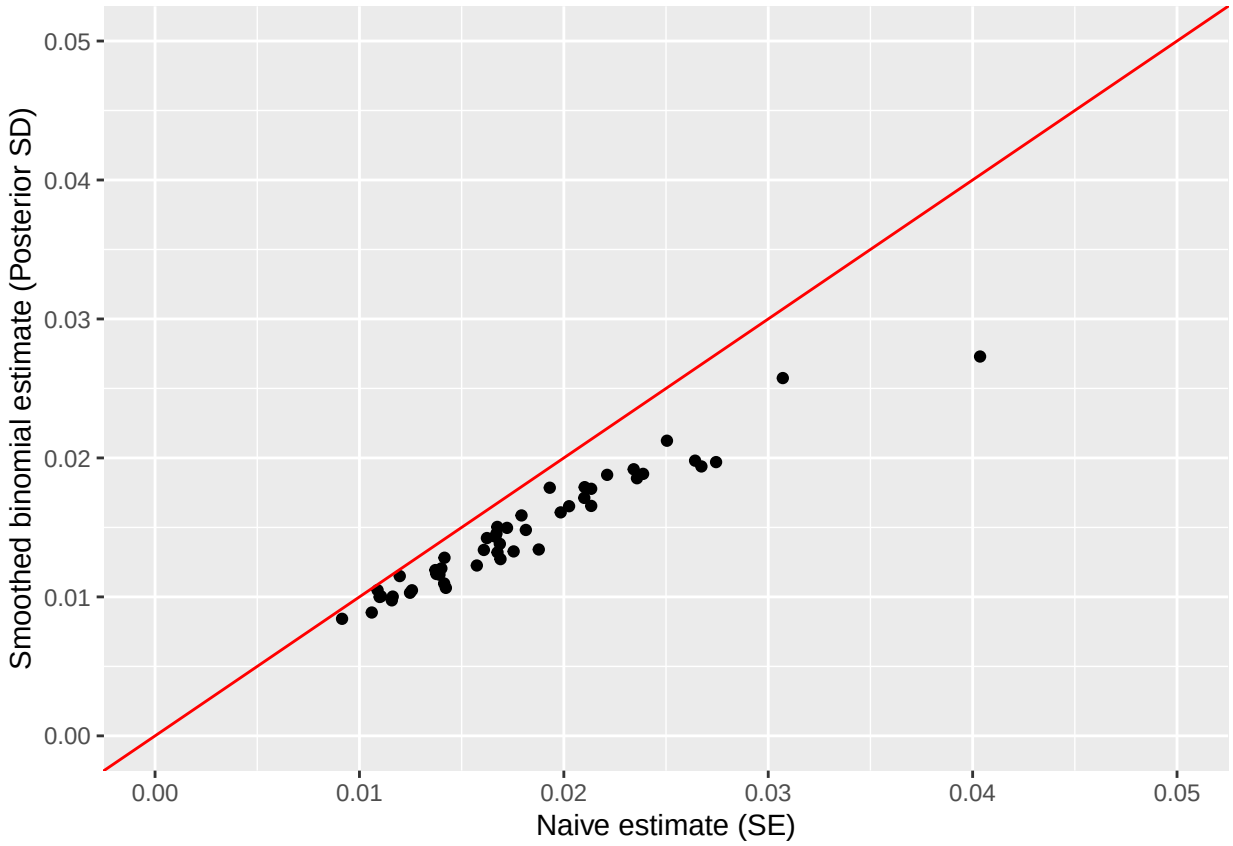
Question 5

```
# plot point estimates against each other
data.frame(smoothed = toplot$median, naive = naive_data$est) %>%
  ggplot(aes(naive, smoothed)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red") +
  coord_fixed() +
  xlab("Naive estimate") +
  ylab("Smoothed binomial estimate")
```



The smoothed binomial estimates are roughly the same as the naive estimates for each region, though we can see that the range of the smoothed binomial estimates is slightly smaller than that of the naive estimates, implying some degree of smoothing has taken place.

```
# plot standard error and posterior sd against each other
data.frame(smoothed = toplot$sd, naive = naive_data$se) %>%
  ggplot(aes(naive, smoothed)) +
  geom_point() +
  xlab("Naive estimate (SE)") +
  ylab("Smoothed binomial estimate (Posterior SD)") +
  xlim (0, 0.05) +
  ylim (0, 0.05) +
  geom_abline(slope = 1, intercept = 0, col = "red")
```

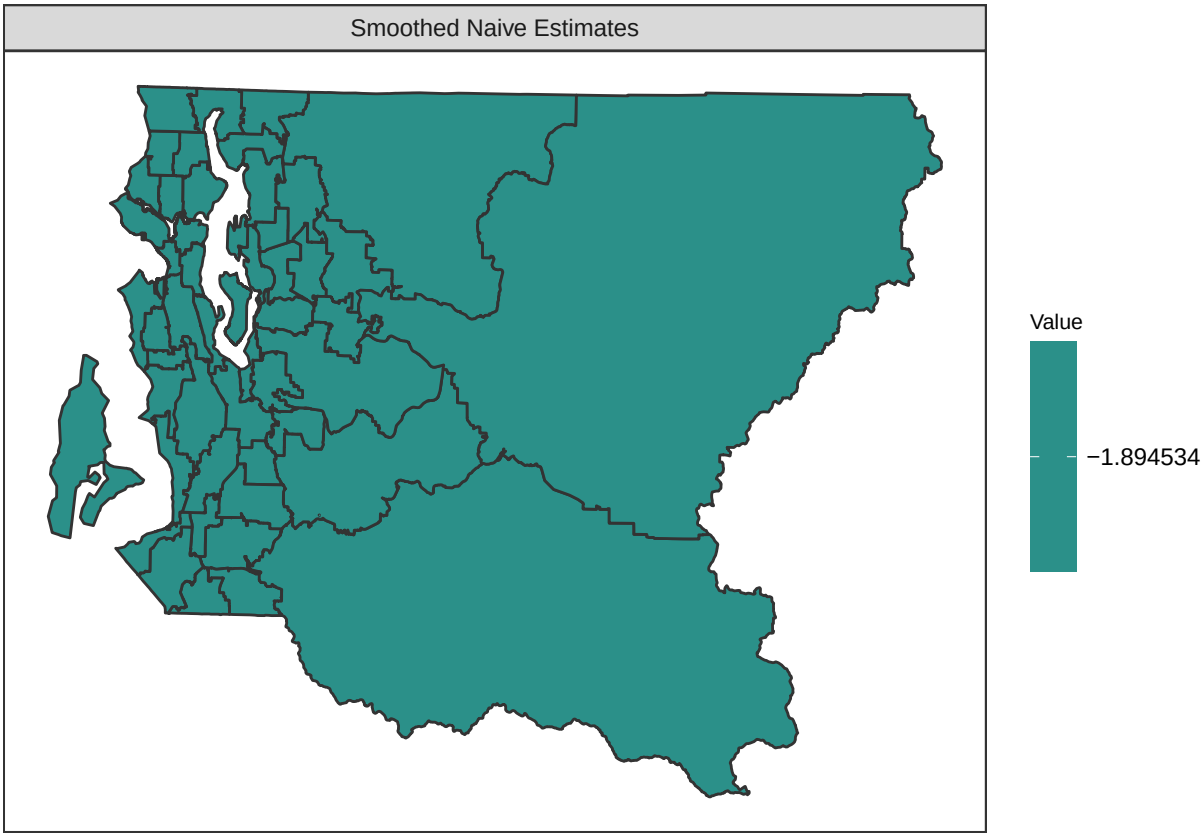


The standard errors of the posterior standard deviations of the smoothed binomial estimates are slightly smaller than the one of the naive estimates. The standard errors of the naive estimates range from around 0.009 to 0.040, whereas the posterior standard deviations of the smoothed binomial estimates range from 0.008 to 0.027. This is again indicative of some degree of smoothing.

Question 6

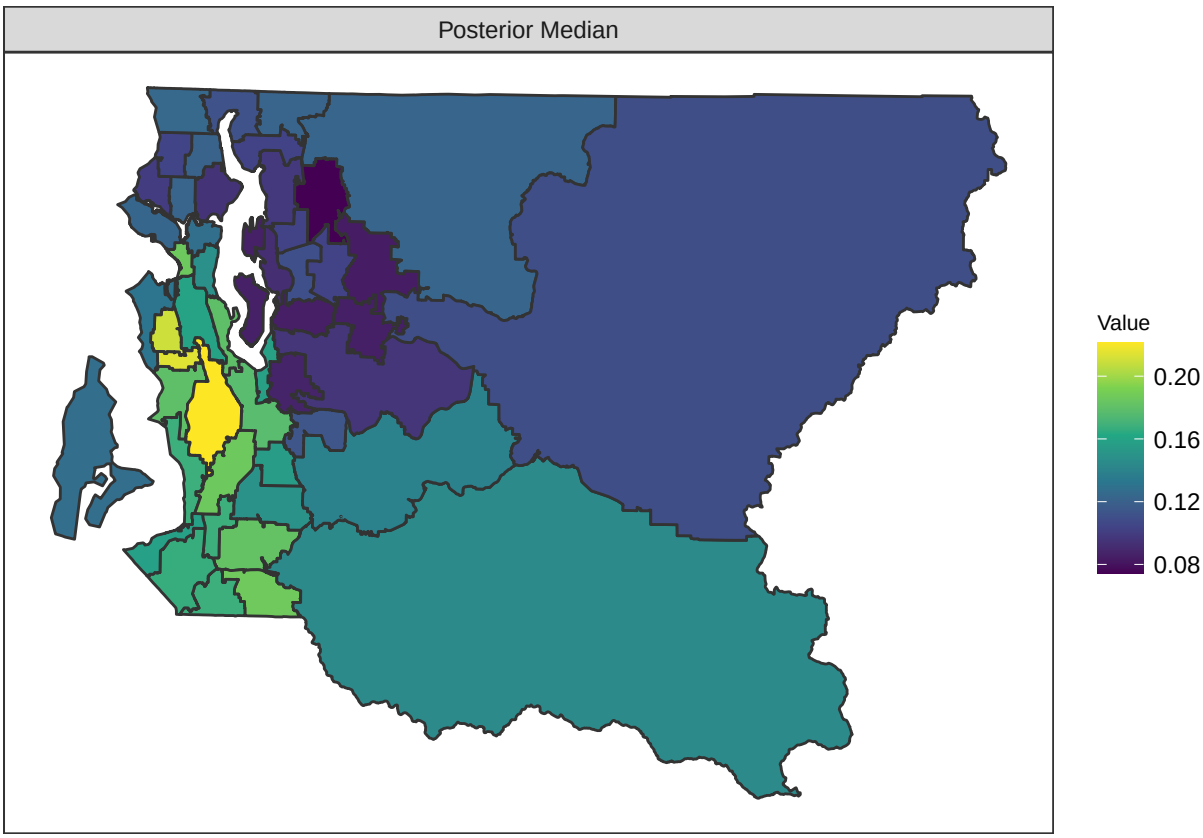
```
# smoothed weighted estimation
FHmodel <- smoothSurvey(data = BRFSS, geo = KingCounty, Amat = mat,
    responseType = "binary", responseVar = "smoker1", strataVar = "strata",
    weightVar = "rwt_llcp", regionVar = "hracode", clusterVar = "~1",
    CI = 0.95)

svysmoothed <- smoothSurvey(data = BRFSS, geo = KingCounty, Amat = mat,
                            responseType = "binary", responseVar = "smoker1",
                            strataVar = "strata",weightVar = "rwt_llcp",
                            regionVar = "hracode", clusterVar = "~1",
                            CI = 0.95)



toplot$fixedpm <- svysmoothed$fit$summary.fixed$`0.5quant`
mapPlot(data = toplot, geo = KingCounty, variables = c("fixedpm"),
        labels = c("Smoothed Naive Estimates"), by.data = "region",
        by.geo = "HRA2010v2_")
```

```
## Using region as id variables
```



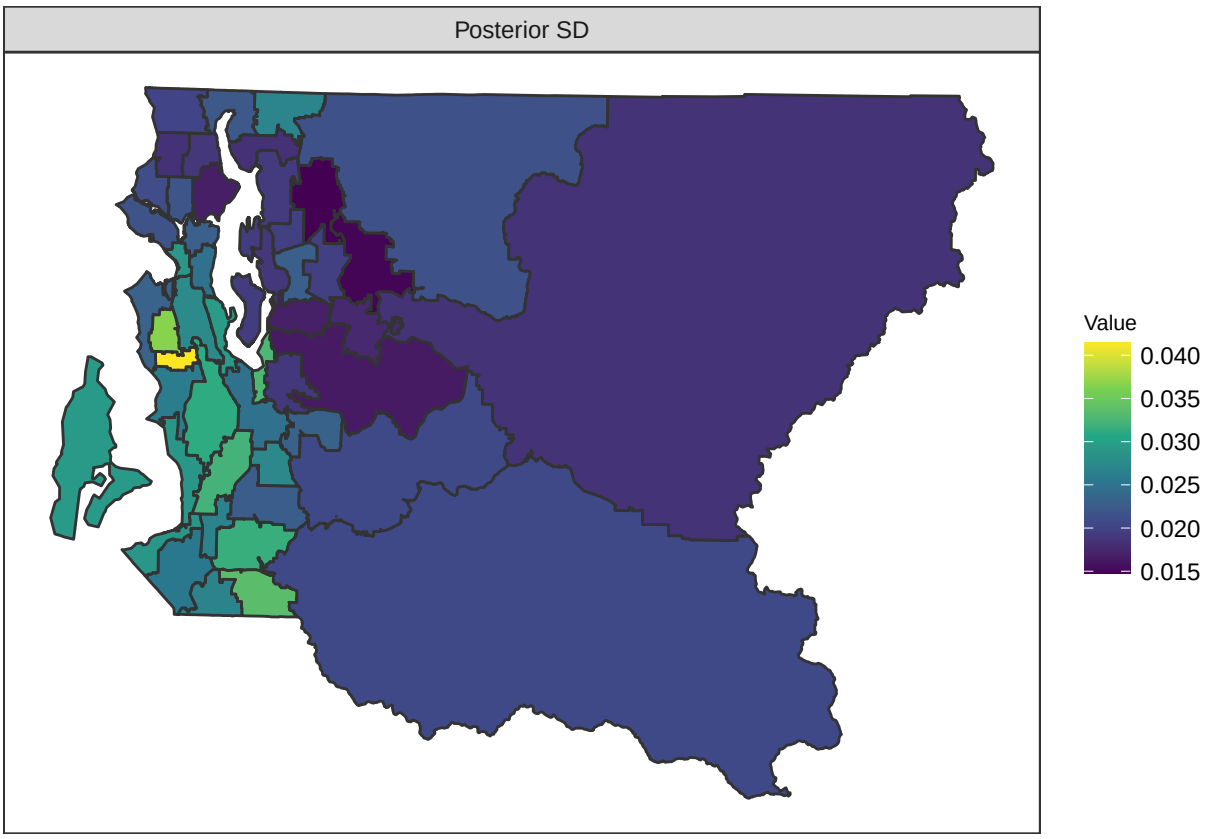

```
# extract posterior medians and map them
toplotFH <- FHmodel$smooth
mapPlot(data = toplotFH, geo = KingCounty, variables = c("median"),
labels = c("Posterior Median"), by.data = "region", by.geo = "HRA2010v2_")
```

```
## Using region as id variables
```



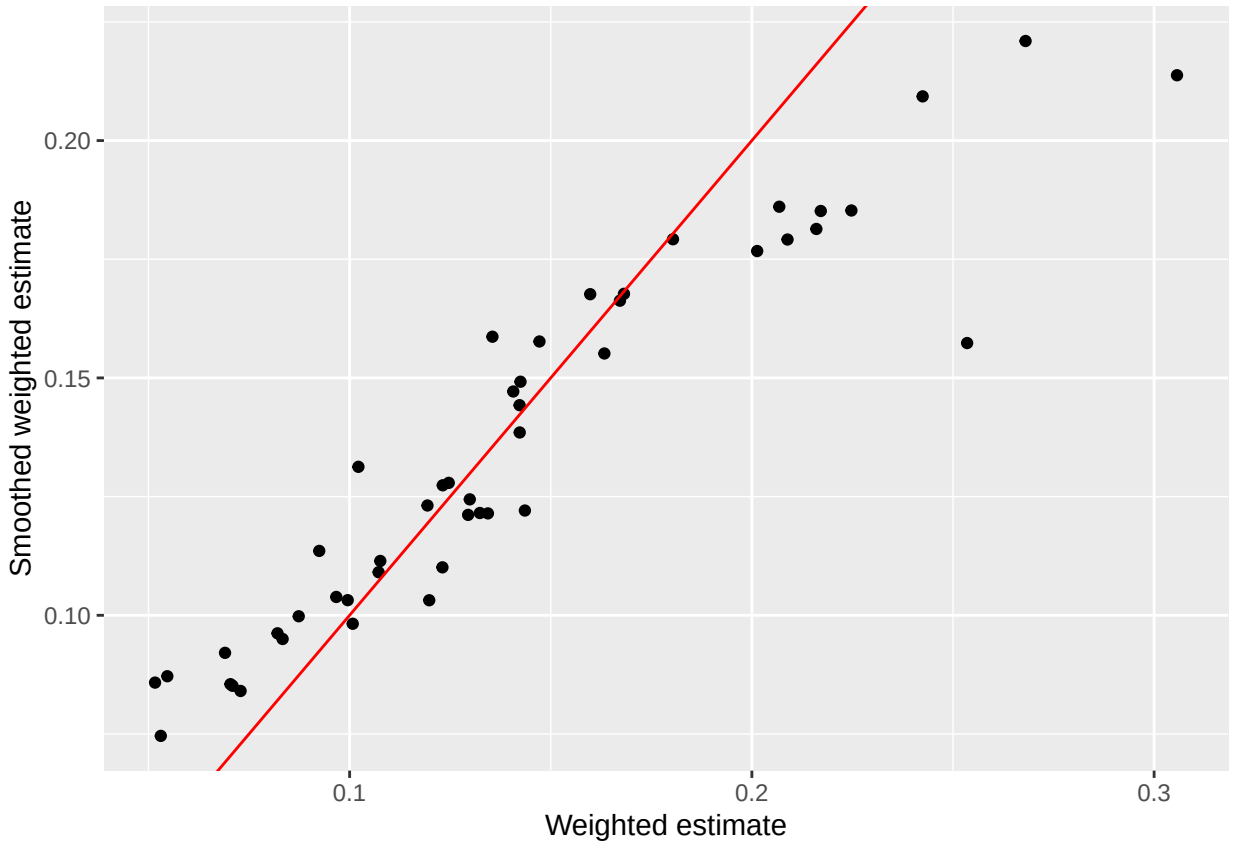

```
# extract posterior standard deviations and map them
toplotFH$sd <- sqrt(toplotFH$var)
mapPlot(data = toplotFH, geo = KingCounty, variables = c("sd"),
labels = c("Posterior SD"), by.data = "region", by.geo = "HRA2010v2_")
```

```
## Using region as id variables
```



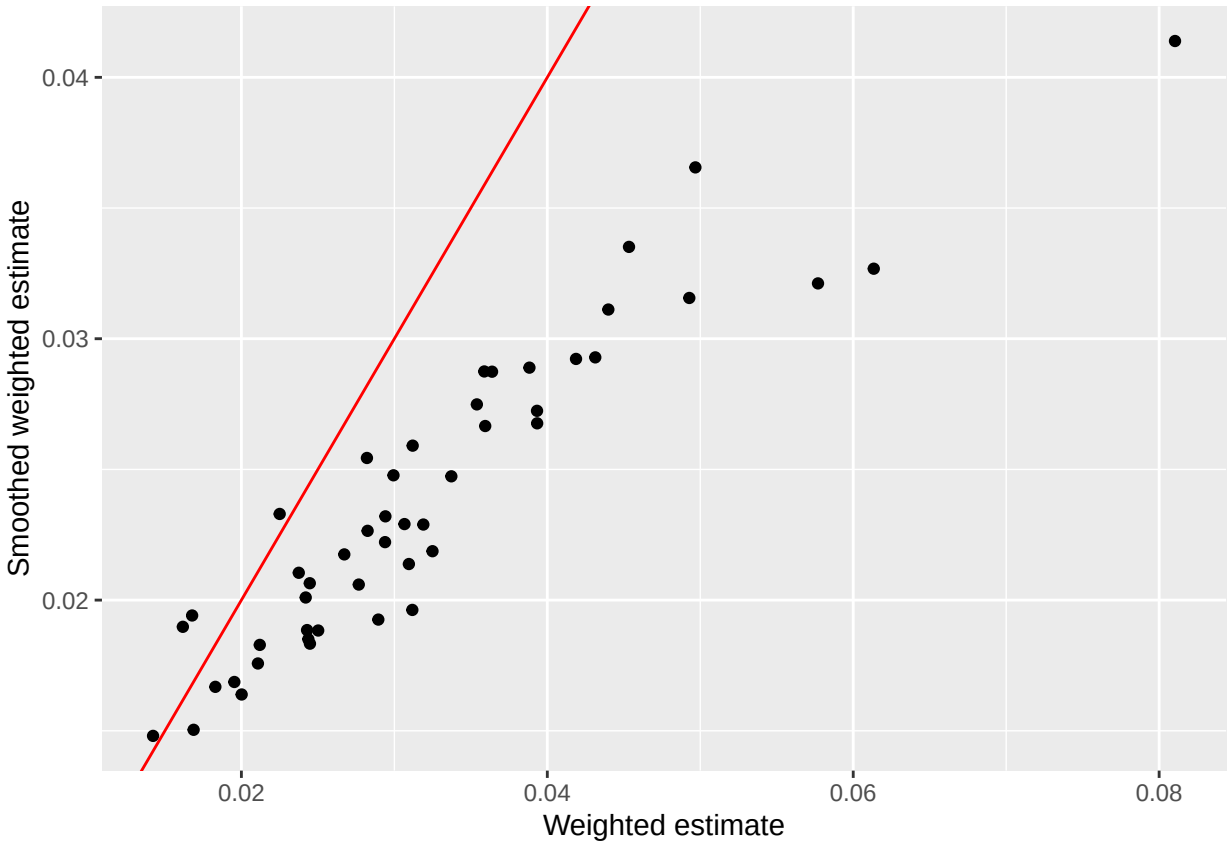
Question 7

```
# plot point estimates against each other
data.frame(smoothed = toplotFH$median, weighted = direct$smoker1) %>%
  ggplot(aes(weighted, smoothed)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red") +
  xlab("Weighted estimate") +
  ylab("Smoothed weighted estimate")
```



The smoothed weighted estimates are smoothed towards the center as expected. Areas with high weighted estimates are shrunk to lower values and areas with low weighted estimates are pulled toward higher values. Smoothed weighted estimates also exhibit lower variablility (a smaller range) compared to weighted estimates.

```
# plot standard error and posterior sd against each other
data.frame(smoothed = toplotFH$sd, weighted = direct$se) %>%
  ggplot(aes(weighted, smoothed)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, col = "red") +
  xlab("Weighted estimate") +
  ylab("Smoothed weighted estimate")
```



The posterior standard deviations from the smoothed weighted model are in general smaller than the posterior standard deviations of the weighted model due to smoothing.

Which of the weighted or the smoothed weighted would you recommend using? Why?

There are two possible justifiable answers:

- I recommend using the smoothed weighted estimates, as the lower variances of the smoothed estimates is desirable (and may be needed) in order to make meaningful scientific conclusions about the proportion of smokers in each area.
- I recommend using the weighted estimates, since they are design consistent, do not require assuming the form of a model, and sample sizes are large enough within each area to provide confidence intervals that are small enough to make meaningful scientific conclusions about the proportion of smokers in each area.

Question 8

In general, smoking prevalence is highest in more Western parts of King County than Eastern parts of King County, with the highest prevalences being in the SeaTac/Tukwila, North Highline, and Delridge HRAs (according to the smoothed weighted model). We can see that HRAs with the lowest smoking prevalence tend to be gathered on the East side of Lake Washington, including Mercer Island and parts of Bellevue. Estimates in the table are reported as percentages.

```
# Summarize the HRA variation in smoking prevalence across King County.

# HRAs with the highest prevalence according to each model? arrange table
```

```r
# according to smoothed weighted estimates

# A table is not necessary for points for this problem, but may be nice to
# help compare HRAs across models
data.frame(region = toplot$region,
           naive = (naive_data$est * 100) %>% round(1),
           weighted = (direct$smoker1 * 100) %>% round(1),
           smoothed = (toplot$median * 100) %>% round(1),
           smoothed_weighted = (toplotFH$median * 100) %>% round(1)) %>%
  dplyr::arrange(smoothed_weighted) %>%
  knitr::kable("simple")
```

region	naive	weighted	smoothed	smoothed_weighted
Redmond	5.7	5.3	6.4	7.5
Sammamish	6.6	7.3	7.1	8.4
Bellevue-South	6.9	7.1	7.4	8.5
Issaquah	8.4	7.0	8.3	8.5
Mercer Isle/Pt Cities	4.3	5.2	6.1	8.6
Renton-East	6.9	5.5	8.7	8.7
Bellevue-West	5.6	6.9	6.9	9.2
NE Seattle	5.7	8.3	6.3	9.5
Newcastle/Four Creeks	9.7	8.2	9.5	9.6
Kirkland	5.7	10.1	6.2	9.8
Ballard	7.3	8.7	7.5	10.0
Bellevue-NE	6.8	12.0	7.0	10.3
Kirkland North	9.6	10.0	8.8	10.3
NW Seattle	9.3	9.7	9.2	10.4
Snoqualmie/North Bend/Skykomish	10.2	10.7	9.9	10.9
Bellevue-Central	8.8	12.3	8.3	11.0
Kenmore/LFP	7.3	10.8	8.0	11.1
Fairwood	9.8	9.2	10.5	11.4
Bothell/Woodinville	9.1	13.4	8.8	12.1
North Seattle	10.6	12.9	9.9	12.1
Bear Creek/Carnation/Duvall	8.0	14.4	8.1	12.2
Fremont/Greenlake	7.4	13.2	7.7	12.2
QA/Magnolia	6.9	11.9	7.5	12.3
Shoreline	11.9	13.0	11.3	12.4
Vashon Island	9.9	12.3	10.1	12.7
Capitol Hill/E.lake	9.1	12.5	9.1	12.8
West Seattle	8.7	10.2	9.7	13.1
Covington/Maple Valley	12.1	14.2	12.0	13.9
Black Diamond/Enumclaw/SE County	12.5	14.2	12.7	14.4
Central Seattle	10.3	14.1	10.8	14.7
Kent-SE	12.6	14.2	12.9	14.9
Kent-East	14.0	16.3	13.4	15.5
Renton-North	15.7	25.4	12.4	15.7
Fed Way-Dash Point/Woodmont	10.9	14.7	12.0	15.8
Beacon/Gtown/S.Park	9.5	13.6	11.5	15.9
Fed Way-Central/Military Rd	15.3	16.7	15.0	16.6
Des Moines/Normandy Park	11.6	16.0	12.3	16.8
East Federal Way	15.1	16.8	14.8	16.8
Renton-South	15.2	20.1	14.3	17.7

region	naive	weighted	smoothed	smoothed_weighted
Burien	14.0	18.0	13.9	17.9
SE Seattle	14.9	20.9	13.8	17.9
Auburn-North	14.8	21.6	14.5	18.1
Downtown	17.6	21.7	15.3	18.5
Kent-West	15.5	22.5	14.6	18.5
Auburn-South	21.3	20.7	19.1	18.6
Delridge	16.4	24.2	15.4	20.9
North Highline	21.9	30.6	17.3	21.4
SeaTac/Tukwila	19.9	26.8	17.7	22.1