SISMID 2021: R Notes Disease Mapping

Jon Wakefield
University of Washington

2021-07-10

SMR Estimates

Scottish lip cancer data

```
library(SpatialEpi)
library(RColorBrewer)
library(ggplot2)
library(ggridges)
library(INLA)
```

Scottish lip cancer data

We will first fit a number of models to the famous Scottish lip cancer data.

We have counts of disease, expected numbers and an area-based covariate (proportion in agriculture, fishing and farming) in each of 56 areas.

```
data(scotland)
Y <- scotland$data$cases
X <- scotland$data$AFF
E <- scotland$data$expected
# Relative risk estimates
smr <- Y/E
summary(E)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.100   4.050   6.300   9.575  10.125  88.700
summary(smr)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.496   1.111   1.522   2.241   6.429
scotland.map <- scotland$spatial.polygon
```

Scottish lip cancer data

The SMRs have a large spread with an increasing trend in the south-north direction.

```
scotd <- scotland$data[, c("county.names",
    "cases", "expected", "AFF")]
scotd$SIR <- scotd$cases/scotd$expected
smap <- scotland$spatial.polygon
sapply(slot(smap, "polygons"), function(x) {
    slot(x, "ID")
})
##  [1] "skye-lochalsh" "banff-buchan"  "caithness"     "berwickshire"
##  [5] "ross-cromarty" "orkney"        "moray"         "shetland"
##  [9] "lochaber"      "gordon"        "western.isles" "sutherland"
## [13] "nairn"         "wigtown"       "NE.fife"       "kincardine"
## [17] "badenoch"      "ettrick"       "inverness"     "roxburgh"
## [21] "angus"         "aberdeen"      "argyll-bute"   "clydesdale"
## [25] "kirkcaldy"     "dunfermline"   "nithsdale"     "east.lothian"
## [29] "perth-kinross" "west.lothian"  "cumnock-doon"  "stewartry"
## [33] "midlothian"    "stirling"      "kyle-carrick"  "inverclyde"
## [37] "cunninghame"   "monklands"     "dumbarton"     "clydebank"
## [41] "renfrew"       "falkirk"       "clackmannan"   "motherwell"
## [45] "edinburgh"     "kilmarnock"    "east.kilbride" "hamilton"
## [49] "glasgow"       "dundee"        "cumbernauld"   "bearsden"
## [53] "eastwood"      "strathkelvin"  "tweeddale"     "annandale"
rownames(scotd) <- scotd$county
smap <- SpatialPolygonsDataFrame(scotland.map,
    scotd, match.ID = TRUE)
```

Scottish lip cancer data

The SMRs have a large spread with an increasing trend in the south-north direction.

```
spplot(smap, zcol = "SIR", col.regions = brewer.pal(9, "Purples"), cuts = 8)
```

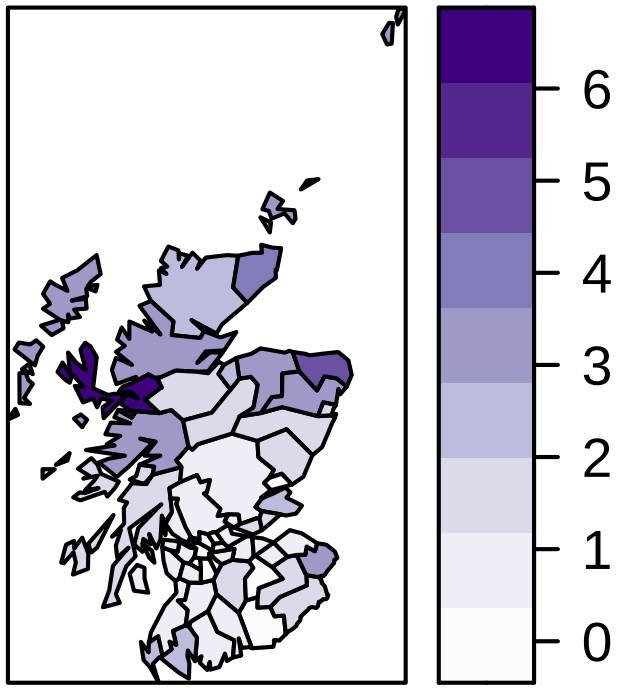


Figure 1: SMRs for Scottish lip cancer data

Scottish lip cancer data

The variance of the estimate in area i is

$$\text{var}(\text{SMR}_i) = \frac{\text{SMR}_i}{E_i},$$

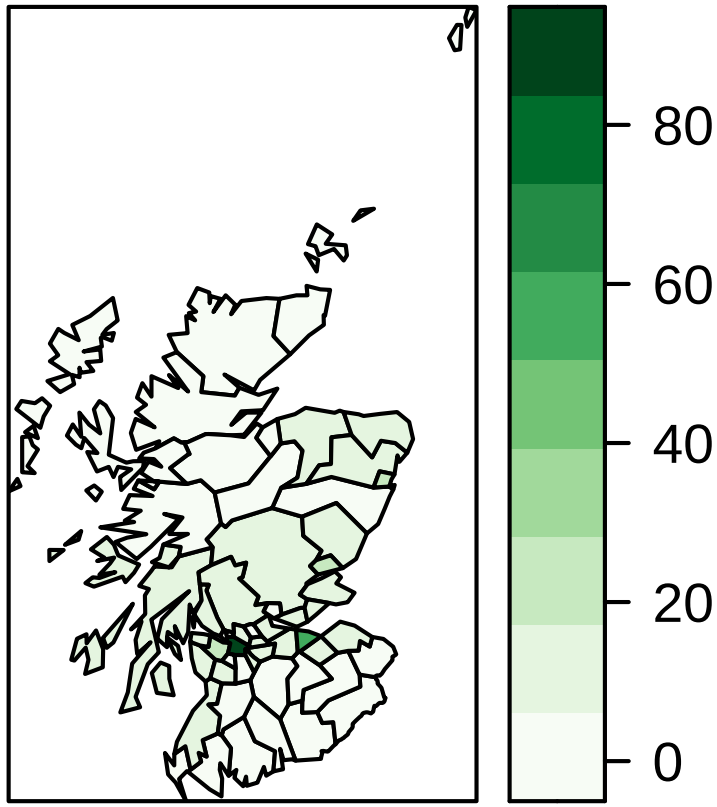
which will be large if E_i is small.

For the Scottish data the expected numbers are highly variable, with range 1.1–88.7.

This variability suggests that there is a good chance that the extreme SMRs are based on small expected numbers (many of the large, sparsely-populated rural areas in the north have high SMRs).

Expected numbers for Scottish lip cancer data

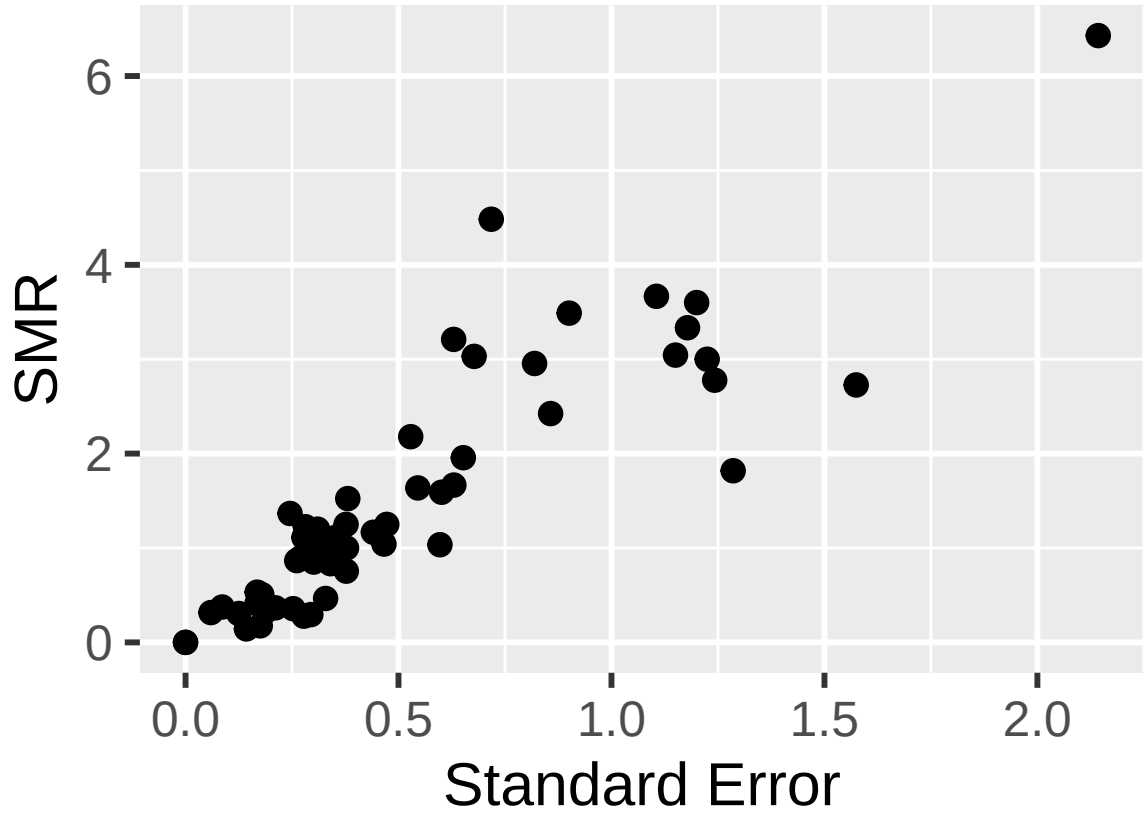
```
spplot(smap, zcol = "expected", col.regions = brewer.pal(9, "Greens"), cuts = 8)
```



SMR for Scottish lip cancer data

The highest SMRs tend to have the largest standard errors.

```
ggplot(data.frame(se = sqrt(smr/E), smr), aes(x = se, y = smr)) +
    geom_point() + labs(y = "SMR", x = "Standard Error")
```



Lognormal Non-Spatial Smoothing Model

Lognormal model

We now consider an alternative lognormal model for the relative risks, but still independent.

A Poisson-lognormal non-spatial random effect model is given by:

$$\begin{aligned} Y_i|\beta_0, e_i &\sim_{ind} \text{Poisson}(E_i e^{\beta_0} e^{e_i}), \\ e_i|\sigma_e^2 &\sim_{iid} \text{N}(0, \sigma_e^2) \end{aligned}$$

where e_i are area-specific random effects that capture the residual or unexplained (log) relative risk of disease in area i, $i = 1, ..., n$.

Note that in INLA the uncertainty in the distribution of the random effect is reported in terms of the precision (the reciprocal of the variance).

Lognormal model

This model gives rise to the posterior distribution;

$$p(\beta_0, \tau_e, e_1, \ldots, e_n|y) = \frac{\prod_{i=1}^{n} \Pr(Y_i|\beta_0, e_i)p(e_i|\tau_e)p(\beta_0)p(\tau_e)}{\Pr(y)}.$$

The full posterior is an $(n+2)$-dimensional distribution and INLA by default produces summaries of the univariate posterior distributions for β_0 and τ_e.

The posteriors on the random effects $p(e_i|y)$ can be extracted, as we will show in subsequent slides.

INLA for lognormal model

We fit the Poisson-Lognormal model for Scotland.

```
# Fit Poisson-lognormal model in INLA:
pcprec <- list(theta=list(prior='pc.prec',param=c(1,.05)))
scotland.fit1 <- inla(Counts ~ 1 + f(Region, model="iid",
  hyper=pcprec),
data=Scotland, family="poisson", E=E,
# Next part calculates fitted values
control.predictor = list(compute = TRUE))
fit1fitted <- scotland.fit1$summary.fitted.values$`0.5quant`
```

Notes on INLA for lognormal model

Note the specification of the penalized complexity prior for the precision $\tau_e = \sigma_e^{-2}$. Here we specify that there is a 5% chance that the standard deviation σ_e is greater than 1. The end of these notes contains a brief description of penalized complexity (PC) priots.

The default prior for β_0 (the intercept) is a zero mean normal with a large standard deviation.

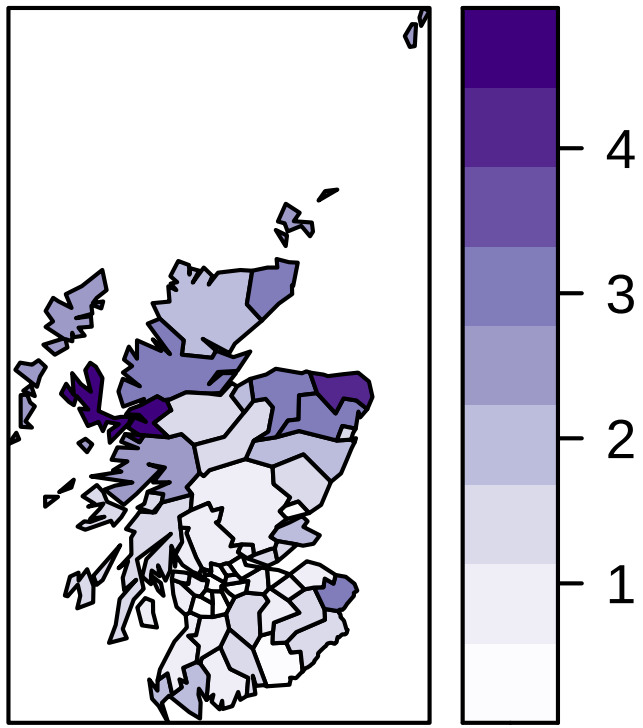
In the `f()` function it is implicit that all random effects are normal.

INLA for lognormal model

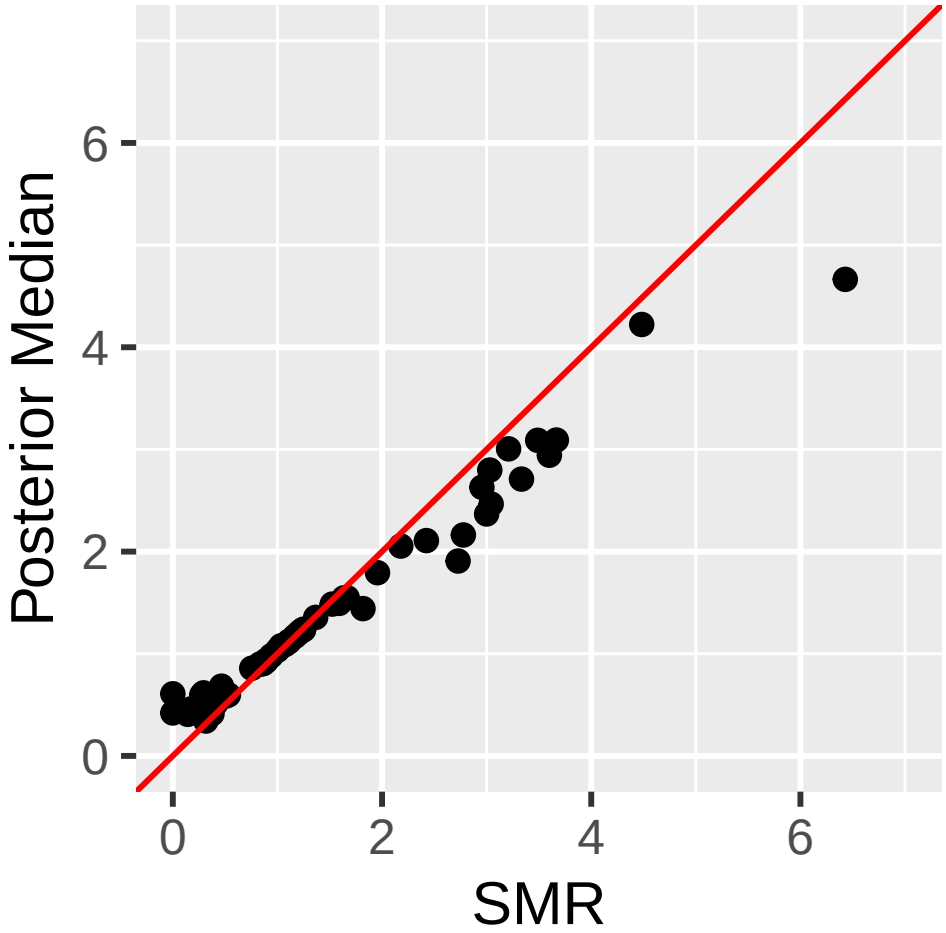
```
names(scotland.fit1)
##  [1] "names.fixed"                "summary.fixed"
##  [3] "marginals.fixed"            "summary.lincomb"
##  [5] "marginals.lincomb"          "size.lincomb"
##  [7] "summary.lincomb.derived"    "marginals.lincomb.derived"
##  [9] "size.lincomb.derived"       "mlik"
## [11] "cpo"                        "po"
## [13] "waic"                       "model.random"
## [15] "summary.random"             "marginals.random"
## [17] "size.random"                "summary.linear.predictor"
## [19] "marginals.linear.predictor" "summary.fitted.values"
## [21] "marginals.fitted.values"    "size.linear.predictor"
## [23] "summary.hyperpar"           "marginals.hyperpar"
## [25] "internal.summary.hyperpar"  "internal.marginals.hyperpar"
## [27] "offset.linear.predictor"    "model.spde2.blc"
## [29] "summary.spde2.blc"          "marginals.spde2.blc"
## [31] "size.spde2.blc"             "model.spde3.blc"
## [33] "summary.spde3.blc"          "marginals.spde3.blc"
## [35] "size.spde3.blc"             "logfile"
## [37] "misc"                       "dic"
## [39] "mode"                       "neffp"
## [41] "joint.hyper"                "nhyper"
## [43] "version"                    "Q"
## [45] "graph"                      "ok"
## [47] "cpu.used"                   "all.hyper"
## [49] ".args"                      "call"
## [51] "model.matrix"
```

INLA for IID lognormal model

```
scotd$fit1fitted <- scotland.fit1$summary.fitted.values$`0.5quant`
smap <- SpatialPolygonsDataFrame(scotland.map,
    scotd, match.ID = TRUE)
spplot(smap, zcol = "fit1fitted", col.regions = brewer.pal(9,
    "Purples"), cuts = 8)
```



```
ggplot(data.frame(pmedian = scotland.fit1$summary.fitted.values$`0.5quant`,
    smr), aes(y = pmedian, x = smr)) + geom_point() + labs(y = "Posterior Median",
    x = "SMR") + geom_abline(intercept = 0, slope = 1, color = "red") +
    xlim(0, 7) + ylim(0, 7)
```

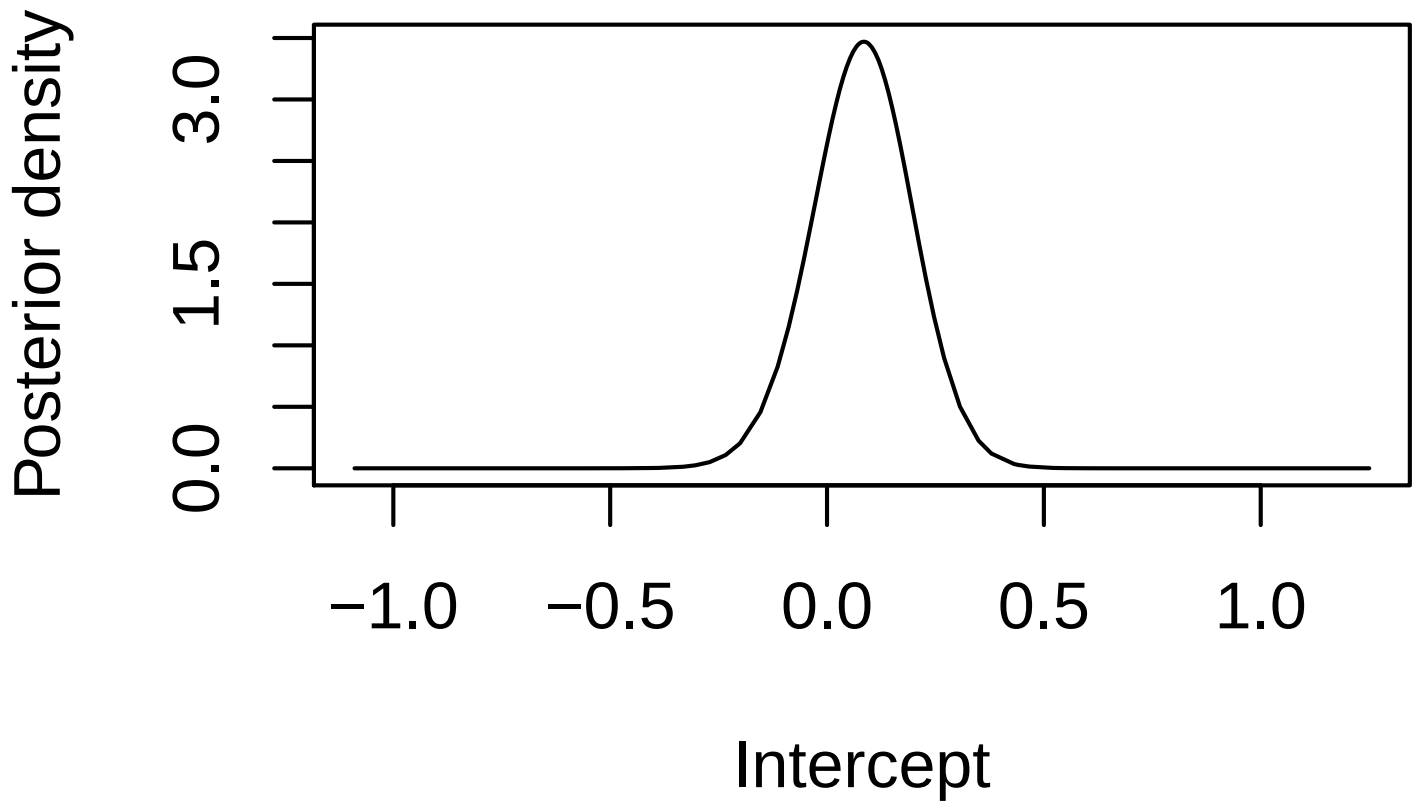


INLA for lognormal model

```
summary(scotland.fit1)
##
## Call:
##    c("inla(formula = Counts ~ 1 + f(Region, model = \"iid\", hyper =
##    pcprec), ", " family = \"poisson\", data = Scotland, E = E,
##    control.predictor = list(compute = TRUE))" )
## Time used:
##     Pre = 6.58, Running = 0.405, Post = 0.412, Total = 7.4
## Fixed effects:
##               mean    sd 0.025quant 0.5quant 0.975quant  mode kld
## (Intercept) 0.081 0.117     -0.154    0.082      0.307 0.085   0
##
## Random effects:
##   Name     Model
##     Region IID model
##
## Model hyperparameters:
##                       mean   sd 0.025quant 0.5quant 0.975quant mode
## Precision for Region 1.80 0.45       1.06     1.75       2.82 1.65
##
## Expected number of effective parameters(stdev): 43.78(2.06)
## Number of equivalent replicates : 1.28
##
## Marginal log-Likelihood:  -185.47
## Posterior marginals for the linear predictor and
##  the fitted values are computed
expbeta0med <- scotland.fit1$summary.fixed[4]  # intercept
sdmed <- 1/sqrt(scotland.fit1$summary.hyperpar[4])  # sd
```

Lognormal model: posterior marginal for the intercept

```
plot(scotland.fit1$marginals.fixed$`(Intercept)`[,
    2] ~ scotland.fit1$marginals.fixed$`(Intercept)`[,
    1], type = "l", xlab = "Intercept", ylab = "Posterior density")
```



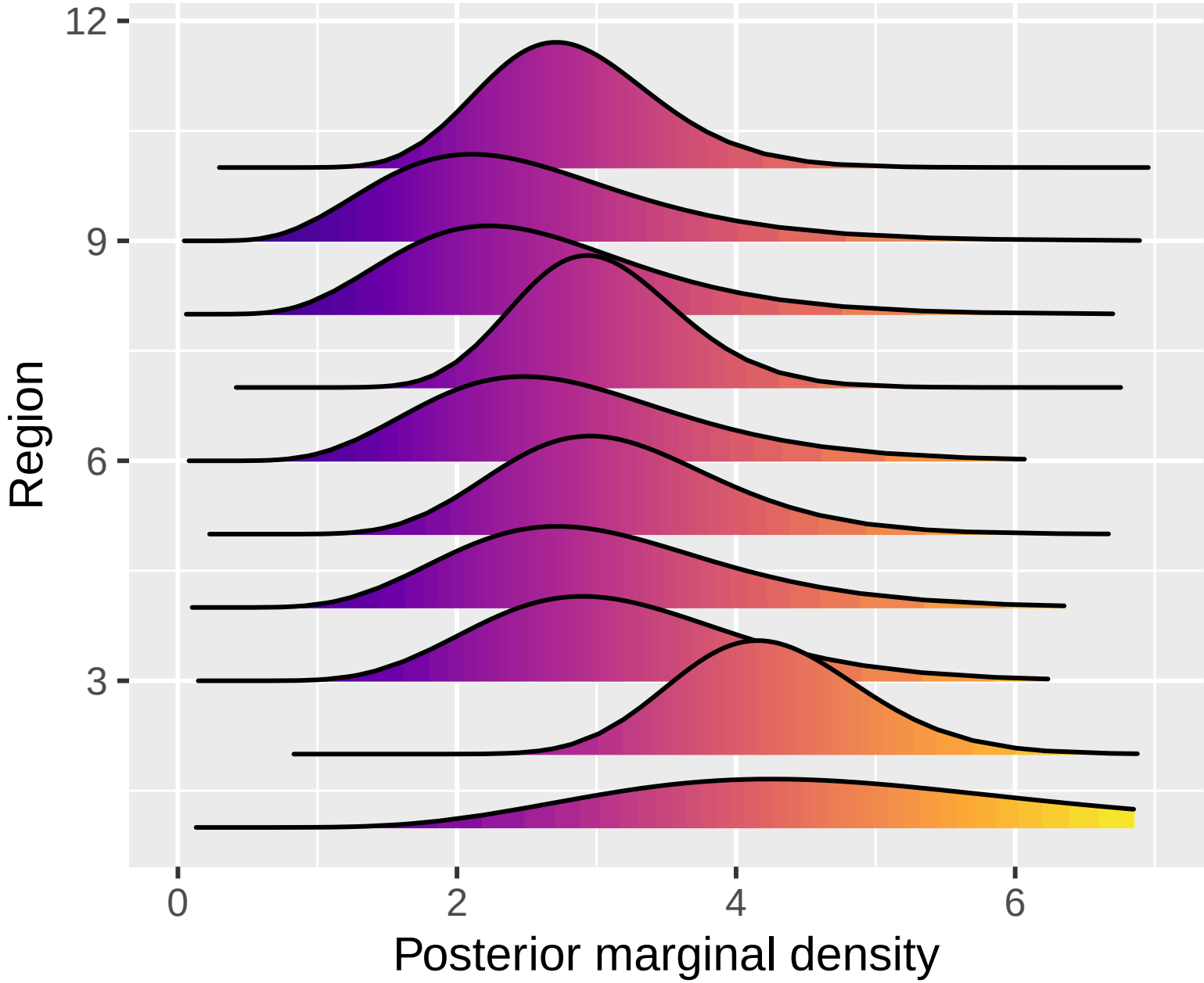
Ridgeplots: posterior marginals for regions

A function to extract a specified marginal for all regions from an INLA model

```
# function to extract the marginal densities and
# make a data frame to plot
extract_marginals_to_plot <- function(marg) {
    posterior_densities <- data.frame()
    for (i in 1:length(marg)) {
        tmp <- data.frame(marg[[i]])
        tmp$Region <- i
        posterior_densities <- rbind(posterior_densities,
            tmp)
    }
    return(posterior_densities)
}
```

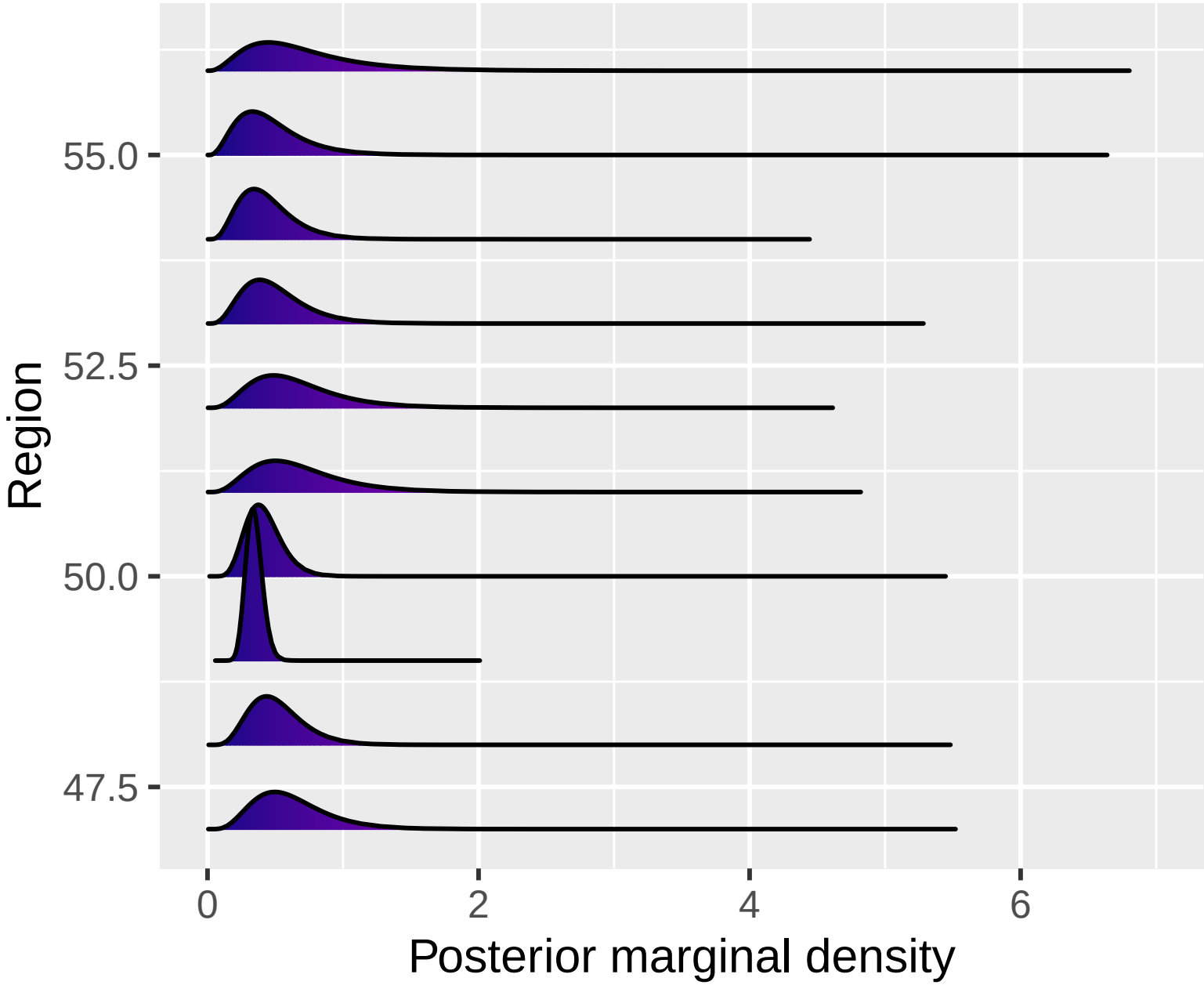
Ridgeplots for marginal posterior RRs in regions 1–10

```
# we now extract the posterior marginal
# distributions of the estimated RRs
marginal_of_interest <- scotland.fit1$marginals.fitted.values
post_dens <- extract_marginals_to_plot(marginal_of_interest)
# we use the ggridges package to plot the marginals
# for first 28 Regions
ggplot(data = post_dens[post_dens$Region <= 10, ],
    aes(x = x, y = Region, height = y, group = Region,
        fill = ..x..)) + geom_density_ridges_gradient(stat = "identity",
    alpha = 0.5) + scale_fill_viridis_c(option = "C") +
    xlab("Posterior marginal density") + xlim(0, 7) +
    theme(legend.position = "none")
```


12
9
6
3
Region
0
2
4
6
Posterior marginal density

Ridgeplots for marginal posterior RRs in regions 47–56

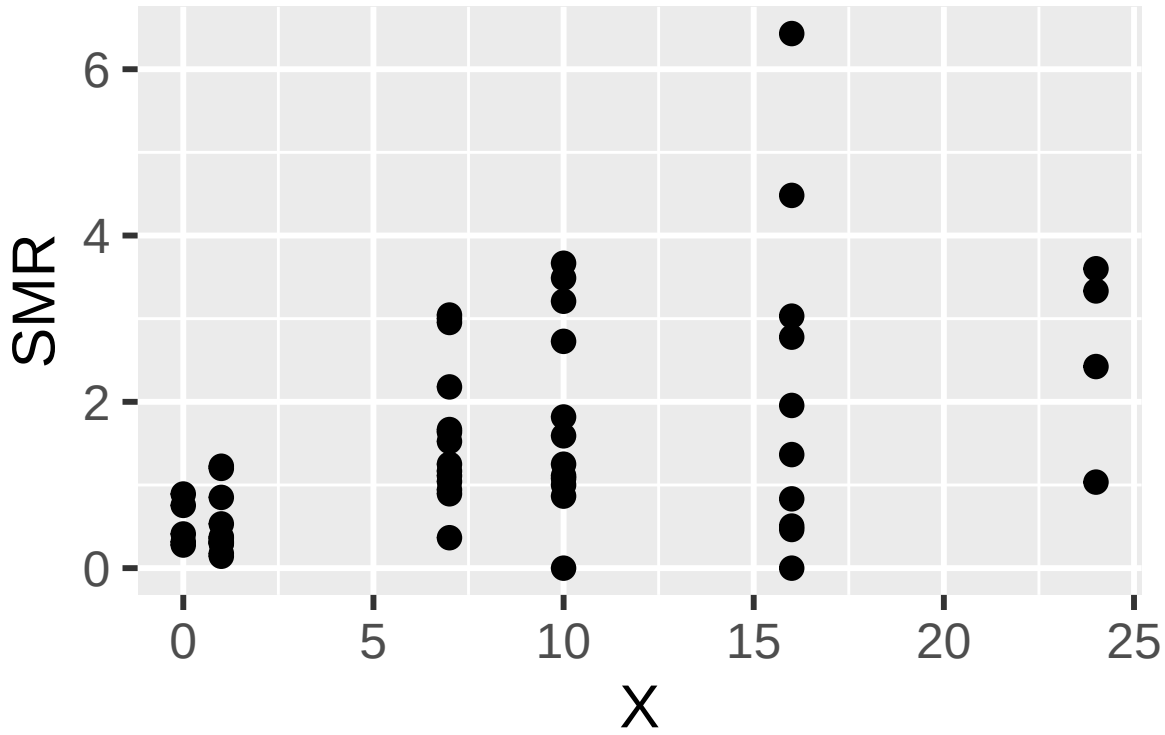
```
# we use the ggridges package to plot the marginals
# for last 10 Regions
ggplot(data = post_dens[post_dens$Region > 46, ], aes(x = x,
    y = Region, height = y, group = Region, fill = ..x..)) +
    geom_density_ridges_gradient(stat = "identity",
        alpha = 0.5) + scale_fill_viridis_c(option = "C") +
    xlab("Posterior marginal density") + xlim(0, 7) +
    theme(legend.position = "none")
```


Region
55.0
52.5
50.0
47.5
0
2
4
6
Posterior marginal density

Add the covariate

We now add AFF, as a sanity check we first plot the SMR versus AFF.

```
ggplot(Scotland, aes(x = X, y = Counts/E)) +
    geom_point() + labs(y = "SMR")
```



Add the covariate

```
modQL <- glm(Scotland$Counts ~ Scotland$X,
    offset = log(Scotland$E), family = "quasipoisson")
coef(modQL)
## (Intercept)  Scotland$X
## -0.54226816  0.07373219
sqrt(diag(vcov(modQL)))
## (Intercept)  Scotland$X
##  0.15418099  0.01320769
```

The estimated RR is $\exp(0.074) = 1.08$, so that an area whose AFF is 1 unit higher has an 8% higher relative risk – not an individual-level association (beware the ecological fallacy!)

Scottish lip cancer

We now fit the three-stage model:

Stage 1: The Likelihood $Y_i|\theta_i \sim \text{Poisson}(E_i\theta_i)$, $i = 1, \ldots, n$ with

$$\log \theta_i = \beta_0 + x_i\beta_1 + e_i$$

where x_i is the AFF in area i.

Stage 2: The random effects (prior distribution) is $e_i|\sigma_e^2 \sim_{iid} N(0, \sigma_e^2)$.

Stage 3: The hyperprior on the hyperparameters $\beta_0, \beta_1, \sigma_e^2$:

$$p(\beta_0, \beta_1, \sigma_e^2) = p(\beta_0)p(\beta_1)p(\sigma_e^2)$$

so that here we have assumed independent priors.

Lognormal non-spatial model with covariates

```
# No spatial effects with covariate
scotland.fit1X <- inla(Counts ~ 1 + X + f(Region, model = "iid",
    hyper = pcprec), data = Scotland, family = "poisson",
    E = E)
scotland.fit1X$summary.fixed[1:5]
##                    mean         sd  0.025quant    0.5quant  0.975quant
## (Intercept) -0.49197093 0.15970410 -0.81195034 -0.49016302 -0.18230136
## X            0.06836979 0.01425663  0.04026998  0.06836384  0.09646857
scotland.fit1X$summary.hyperpar[1:5]
##                             mean        sd 0.025quant 0.5quant 0.975quant
## Precision for Region 2.936798 0.8387875   1.627968 2.823839   4.895032
```

Lognormal non-spatial model with covariates: inference

If we are interested in the association with the AFF variable we can examine the posterior summaries, on the original (to give a log RR) or exponentiated (to give a RR) scale.

From these summaries we might extract the posterior median as a point estimate, or take the 2.5% and 97.5% points as a 95% credible interval.

```
scotland.fit1X$summary.fixed[2, 1:5]
##         mean         sd 0.025quant   0.5quant 0.975quant
## X 0.06836979 0.01425663 0.04026998 0.06836384 0.09646857
exp(scotland.fit1X$summary.fixed[2, 1:5])
##       mean       sd 0.025quant 0.5quant 0.975quant
## X 1.070761 1.014359   1.041092 1.070755   1.101275
```

Parameter interpretation

```
scotland.fit1X$summary.fixed[1:5]
##                    mean         sd  0.025quant    0.5quant  0.975quant
## (Intercept) -0.49197093 0.15970410 -0.81195034 -0.49016302 -0.18230136
## X            0.06836979 0.01425663  0.04026998  0.06836384  0.09646857
```

The posterior mean for the intercept is $E[\beta_0|y] = -0.49$.

The posterior median for the relative risk associated with a 1 unit increase in X is median$(\exp(\beta_1)|y) = \exp(0.068) = 1.07$. This latter calculation exploits the fact that we can transform quantiles[1]

Similarly a 95% credible interval for the relative risk $\exp(\beta_1)$ is

$$[\ \exp(0.040), \exp(0.096)\ ] = [\ 1.04, 1.10\ ].$$

Examination of such intervals is a common way of determining whether the association is "significant" – here we have strong evidence that the relative risk associated with AFF is significant.

[1] unlike means since, for example, $E[\exp(\beta_1)|y] \neq \exp(E[\beta_1|y])$.

Scottish Lip Cancer: Parameter Interpretation

```
scotland.fit1X$summary.fixed[1:5]
##                    mean         sd  0.025quant    0.5quant  0.975quant
## (Intercept) -0.49197093 0.15970410 -0.81195034 -0.49016302 -0.18230136
## X            0.06836979 0.01425663  0.04026998  0.06836384  0.09646857
```

The posterior median of σ_e is $1/\sqrt{2.8} = 0.582$ and a 95% interval is

$$[1/\sqrt{5.13}, 1/\sqrt{1.70}] = [0.44, 0.766].$$

A more interpretable quantity is an interval on the residual relative risk (RRR). The latter follow a lognormal distribution LogNormal$(0, \sigma_e^2)$ so a 95% interval is $\exp(\pm 1.96 \times \sigma_e)$.

Scottish Lip Cancer: Parameter Interpretation

A posterior median of a 95% RRR interval is

$$\begin{aligned}&[\exp(-1.96 \times \text{median}(\sigma_e)), \exp(1.96 \times \text{median}(\sigma_e)] \\ = &[\exp(-1.96 \times 0.582), \exp(1.96 \times 0.582)] = [0.320, 3.13]\end{aligned}$$

which is quite wide.

A more in depth analysis would examine the prior sensitivity to the prior on τ_e.

Variances are in general more difficult to estimate than regression coefficients so there is often sensitivity (unless the number of areas is very large).

Lognormal Spatial Smoothing Model

Lognormal spatial model with one covariate

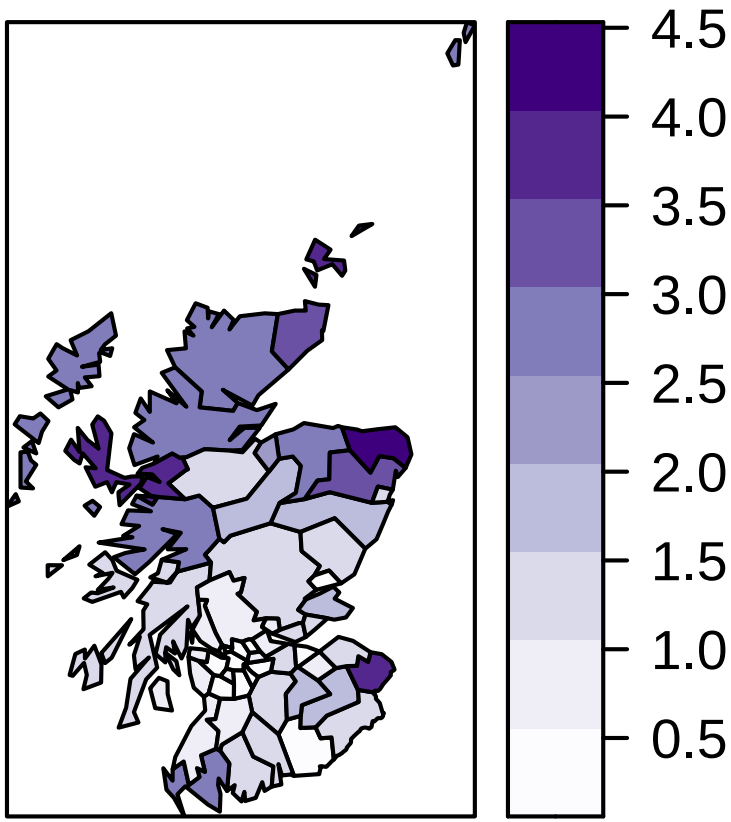
We now add spatial (ICAR) random effects to the model. We parameterize in terms of total variance and proportion that is spatial. We place a penalized complexity prior on these two parameters.

We need a graph file containing the neighbors.

```
# Spatial effects with covariate
download.file("http://faculty.washington.edu/jonno/SISMIDmaterial/scotland.graph",
    destfile = "R-examples/scotland.graph")
formula <- Counts ~ 1 + X + f(Region, model = "bym2",
    graph = "R-examples/scotland.graph", scale.model = T,
    constr = T, hyper = list(phi = list(prior = "pc",
        param = c(0.5, 0.5), initial = 1), prec = list(prior = "pc.prec",
        param = c(0.5/0.31, 0.01), initial = 5)))
scotland.fit2 <- inla(formula, data = Scotland, family = "poisson",
    E = E, control.predictor = list(compute = TRUE),
    control.compute = list(config = TRUE))
```

INLA for spatial lognormal model

```
scotd$fit2fitted <- scotland.fit2$summary.fitted.values$`0.5quant`
smap <- SpatialPolygonsDataFrame(scotland.map,
    scotd, match.ID = TRUE)
spplot(smap, zcol = "fit2fitted", col.regions = brewer.pal(9,
    "Purples"), cuts = 8)
```



Lognormal spatial model with covariates

```
scotland.fit2$summary.hyperpar[, 1:5]
##                              mean          sd 0.025quant  0.5quant 0.975quant
## Precision for Region 4.6384096 1.43531057  2.4193529 4.4405609  8.0152864
## Phi for Region       0.9412494 0.07083046  0.7398139 0.9673838  0.9991849
```

The posterior median of the total standard deviation (on the log relative risk scale) is $1/\sqrt{4.45} = 0.47$.

The posterior median for the proportion of the residual variation that is spatial is 0.96.

Lognormal spatial model with covariates

Now we provide maps of the non-spatial and spatial random effects.

Estimates of residual relative risk (posterior medians), of the non-spatial e^{e_i} and the spatial contributions e^{S_i}.

The BYM2 formulation for the random effect is $b_i = S_i + e_i$ where S_i is spatial and e_i is IID. INLA stores b_i (the first 56 rows) and S_i (the next 56 rows) and so we find the non-spatial via $e_i = b_i - S_i$.

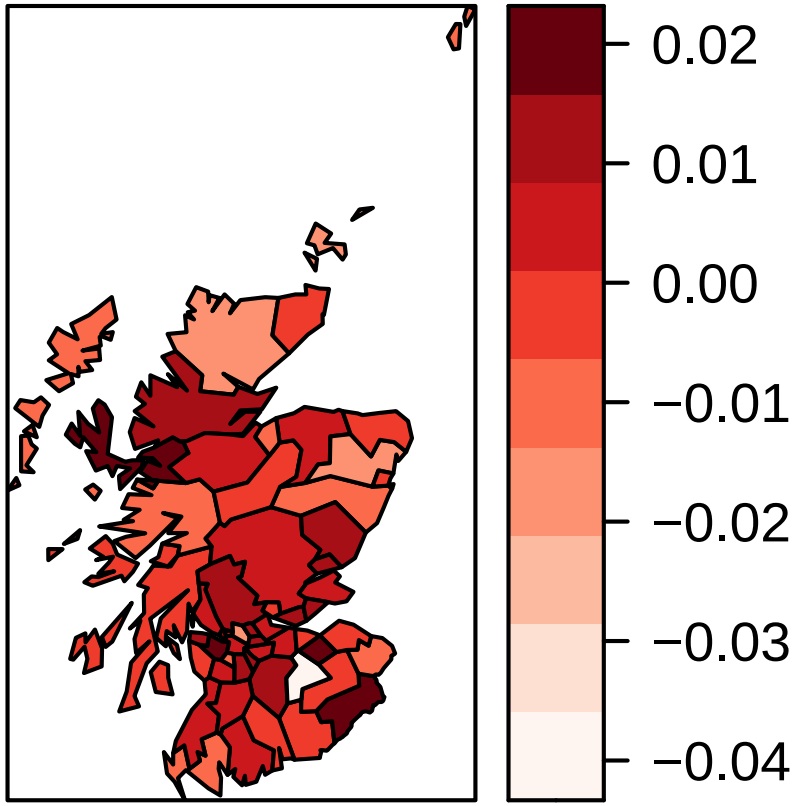
Note the differences in the scales: the spatial random effects dominate here.

```
samp <- inla.posterior.sample(n = 1000, scotland.fit2)
samp_mat <- matrix(0, nrow = 1000, ncol = 2)
for (i in 1:1000) {
    samp_mat[i, ] <- samp[[i]]$hyperpar[1:2]
}
scale_region <- mean(sqrt(samp_mat[, 2])/sqrt(samp_mat[,
    1]))
```

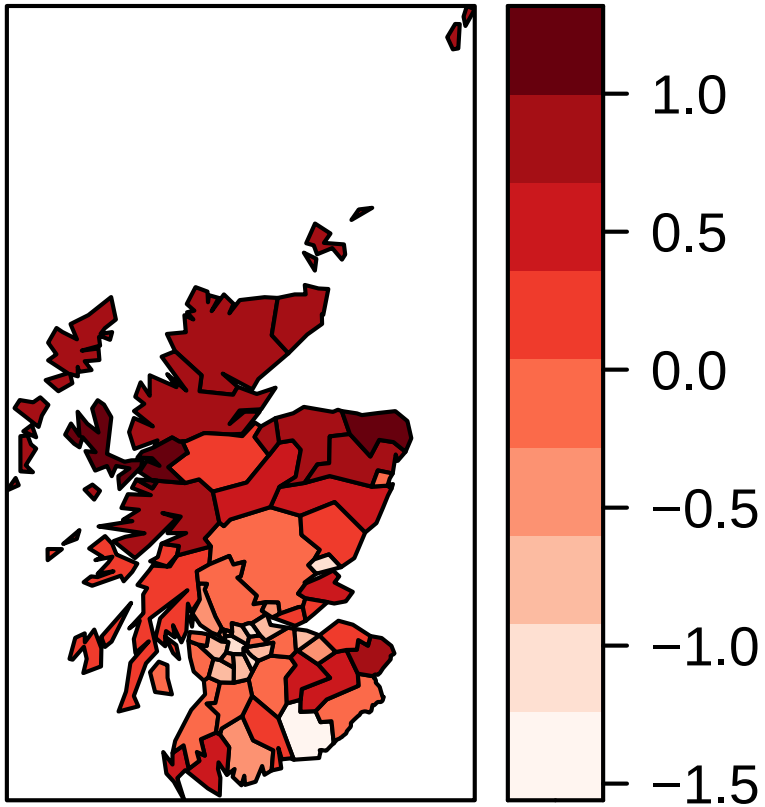
Lognormal spatial model with covariates: non-spatial random effects

```
# obtain RE estimates
N <- 56
struct <- scotland.fit2$summary.random[[1]]$mean[(N +
    1):(N * 2)]
combined <- scotland.fit2$summary.random[[1]]$mean[1:N]
struct <- struct * scale_region
iid <- combined - struct
REsnonspat <- iid
REsspat <- struct
scotd$REsnonspat <- iid
scotd$REsspat <- struct
```

Non-spatial random effects



Spatial random effects



Spatial model: confounding by location

The command `plot(scotland.fit2)` provides plots of: marginal posterior distributions of β_0, β_1, σ_e^{-2}, σ_S^{-2} and summaries of the random effects e_i, S_i and the linear predictors and fitted values, all by area.

Note that the posterior mean estimate of β_1 associated with AFF goes from $0.068 \rightarrow 0.026$ when moving from the non-spatial to spatial model.

This is known as confounding by location.

The model attributes spatial variability in risk to either the covariate or to the spatial random effects.

Scotland

The posterior median estimate of σ_e decreases from $1/\sqrt{2.9475} = 0.58$ to $1/\sqrt{94.986} = 0.10$ when the spatial random effect is added.

The posterior median estimate of σ_s is $1/\sqrt{1.125} = 0.94$ but, as already noted, this value is not directly comparable to the estimate of σ_e.

However, the scales on the figures shows that the spatial component dominates for these data.

A rough estimate of the standard deviation of the spatial component can be determined by empirically calculating the standard deviation of the random effect estimates $\widehat{S}_i$.

A more complete analysis would address the sensitivity to the prior specifications on σ_e and σ_s.

Some Detail

INLA Graph File

The code below creates a neighborhood filefor INLA that looks like:

39

1 4 11 13 22 38 2 2 12 38 3 5 11 13 20 36 39 4 6 9 17 19 24 29 31

...

38 7 1 2 7 11 12 22 32

39 8 3 13 17 19 20 21 27 30

Creating an INLA graph file from a shapefile

```
library(rgdal)  # for readOGR
library(spdep)  # for poly2nb and nb2inla
countymap = readOGR(dsn = "R-examples/wacounty.shp",
    layer = "wacounty")
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/jonno/Dropbox/2020-SISMID/2021-Lectures/2021-SISMID-R-SESSIONS/R-
## with 39 features
## It has 6 fields
nb.map <- poly2nb(countymap)
nb2INLA("wacounty.graph", nb.map)
```

PC prior details

For a precision in the model $x|\tau \sim N(0, 1/\tau)$, the PC prior is obtained via the following rationale:

- The prior on the sd is exponential with rate λ, which we need to specify
- The exponential leads to a type-2 Gumbel on the precision (change of variables)
- Hence we have the model:

$$\begin{aligned} x|\tau &\sim N(0, 1/\tau) \\ \tau &\sim \text{Gumbel}(\lambda) \end{aligned}$$

- If we integrate out τ, we can find the marginal sd of x
- For more details see Simpson et al (2017, p. 9, top of right column) and Bakka et al (2018).