

# 2022 555 Disease Mapping R Notes

Jon Wakefield  
Departments of Biostatistics and Statistics  
University of Washington

2023-01-09

## Scottish lip cancer data

In these notes we will analyze the famous Scottish lip cancer data discussed in lectures. We will fit both non-spatial and spatial random effects smoothing models.

We will also discuss some other topics including how to deal with:

- missing observations, and
- censored observations

We first load some libraries. Note that INLA has a non-standard installation, see [INLA DOWNLOAD](#)

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

We will fit a number of models to the Scottish lip cancer data, but first access the data.

In area  $i$ , let  $Y_i$  and  $E_i$  represent the disease count and expected count.

An initial summary is the Standardized Morbidity Ratio (SMR), which for area  $i$  is

$$\text{SMR}_i = \frac{Y_i}{E_i},$$

for  $i = 1, \dots, 56$ .

We also have an area-based covariate  $X_i$  (proportion in agriculture, fishing and farming) in each of the  $i = 1, \dots, 56$  areas.

```
data(scotland)
names(scotland)
## [1] "geo" "data" "spatial.polygon" "polygon"
names(scotland$data)
## [1] "county.names" "cases" "expected" "AFF"
head(scotland$data)
##   county.names cases expected AFF
## 1 skye-lochalsh    9    1.4 0.16
## 2 banff-buchan   39    8.7 0.16
## 3   caithness   11    3.0 0.10
## 4 berwickshire    9    2.5 0.24
## 5 ross-cromarty  15    4.3 0.10
## 6      orkney    8    2.4 0.24
```

The following is taken from Section 6.2 of Moraga (2020).

We form a data frame `scotdata` containing key variables, and add the SMRs.

```
scotdata <- scotland$data[,c("county.names", "cases", "expected", "AFF")]
scotdata$SMR <- scotdata$cases/scotdata$expected
smap <- scotland$spatial.polygon
```

We use the `supply()` function to see that the polygon ID slot corresponds to the county names.

Then create a polygon `SpatialPolygonsDataFrame`, which allows mapping.

```
supply(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"
```

We now create a `SpatialPolygonsDataFrame` by combining `smap` and `scotdata`.

We first create row names for `scotdata` by setting equal to the county names, and then we match on these row names (`match.ID=TRUE`).

```
rownames(scotdata) <- scotdata$county
smap <- SpatialPolygonsDataFrame(smap, scotdata, match.ID = TRUE)
```

We can look at the first part of the spatial data frame:

```
head(smap@data)
##      county.names cases expected AFF      SMR
## skye-lochalsh skye-lochalsh    9    1.4 0.16 6.428571
## banff-buchan  banff-buchan   39    8.7 0.16 4.482759
## caithness     caithness    11    3.0 0.10 3.666667
## berwickshire  berwickshire    9    2.5 0.24 3.600000
## ross-cromarty ross-cromarty   15    4.3 0.10 3.488372
## orkney        orkney         8    2.4 0.24 3.333333
```

Now for the map.

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

Observations:

- The SMRs have a large spread, but how much does this reflect sampling variation, rather than true variation?
- There is also increasing trend in the south-north direction.

The variance of the estimate in area  $i$  is

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

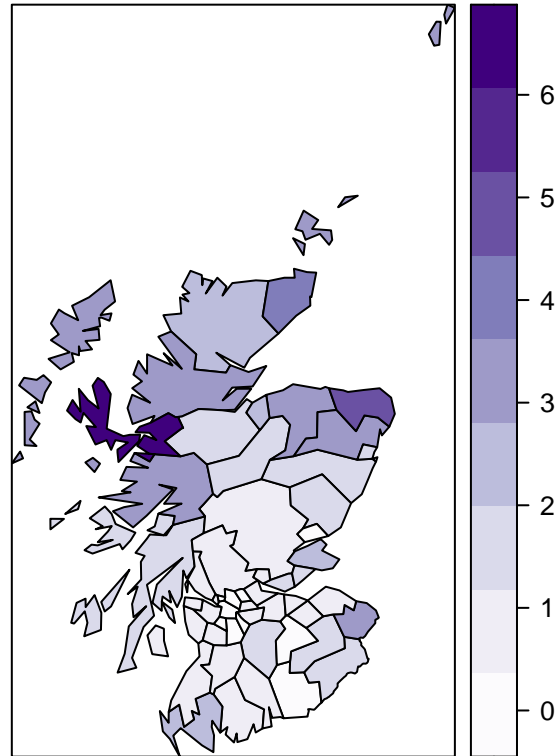


Figure 1: SMRs for Scottish lip cancer data

which will be large if  $E_i$  is small.

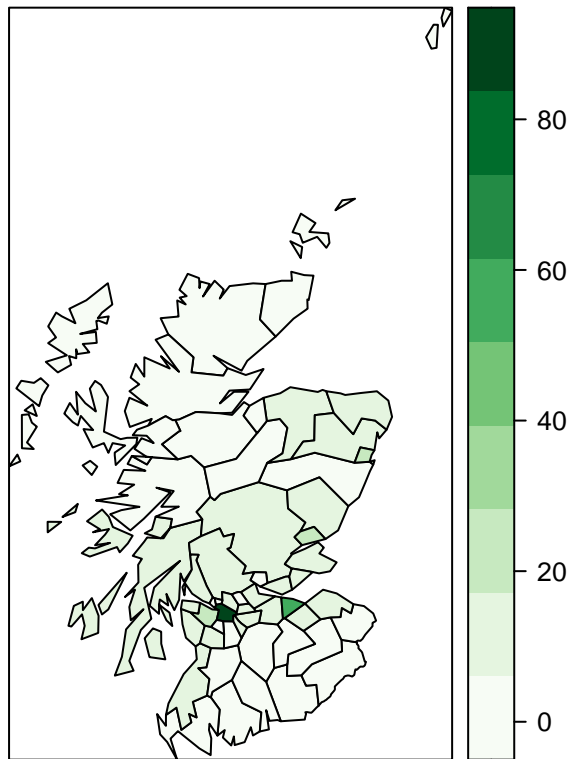
```
summary(smap@data$expected)
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.100  4.050   6.300   9.575 10.125  88.700
```

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).

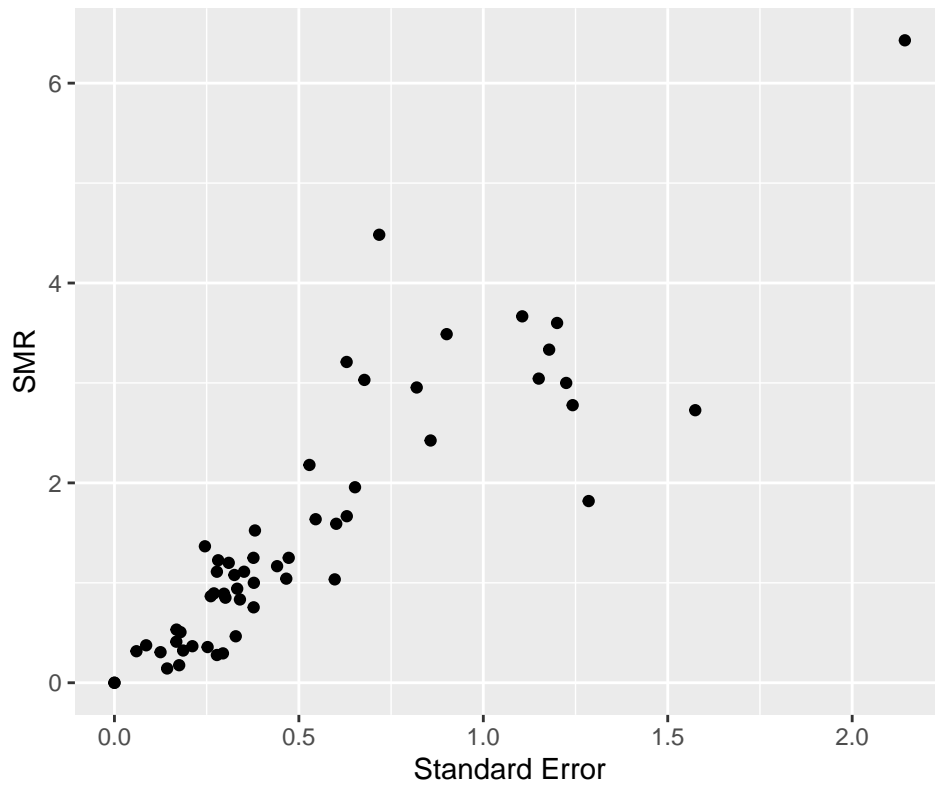
We next map the expected numbers for Scottish lip cancer data

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



The highest SMRs tend to have the largest standard errors.

```
ggplot(data.frame(se=sqrt(scotdata$SMR/scotdata$expected),SMR=scotdata$SMR),aes(x=se,y=SMR)) + geom_point()
```



## SMR interval estimates

Let  $\theta_i = \text{SMR}_i$ .

We obtain an interval estimate for  $\alpha_i = \log \theta_i$  (since the normality of the estimator is likely to be better on this scale) and then transform.

Via the delta method

$$\widehat{\text{var}}(\hat{\alpha}_i) = \widehat{\text{var}}(\hat{\theta}_i) |J|^2$$

where  $J = \frac{d\alpha_i}{d\theta_i} = \exp(-\alpha_i)$  and  $\widehat{\text{var}}(\hat{\theta}_i) = \hat{\theta}_i / E_i$ .

We obtain:

$$\widehat{\text{var}}(\hat{\alpha}_i) = [E_i \exp(\hat{\alpha}_i)]^{-1},$$

to give a 95% confidence interval for  $\theta_i$  of

$$\exp \left( \hat{\alpha}_i \pm 1.96 \times \sqrt{\widehat{\text{var}}(\hat{\alpha}_i)} \right).$$

## SMR estimates when $Y_i = 0$

When  $Y_i = 0$ , we obtain an SMR of 0, and (more worryingly) a standard error of zero.

In this case, we carry out an adjustment and set  $Y_i^* = Y_i + 0.5$  and  $E_i^* = E_i + 0.5$  to give the estimator

$$\theta_i^* = \text{SMR}_i^* = Y_i^* / E_i^*,$$

with  $\widehat{\text{var}}(\hat{\theta}_i^*) = \hat{\theta}_i^* / E_i^*$ .

Also let  $\alpha_i^* = \log \theta_i^*$ .

We obtain:

$$\widehat{\text{var}}(\hat{\alpha}_i^*) = (E_i^* \exp(\hat{\alpha}_i^*))^{-1},$$

to give a 95% confidence interval of

$$\exp \left( \hat{\alpha}_i^* \pm 1.96 \times \sqrt{\widehat{\text{var}}(\hat{\alpha}_i^*)} \right).$$

## SMR interval estimates

The addition of 0.5 is somewhat ad hoc but corresponds to a  $\text{Ga}(0.5, 0.5)$  prior on the relative risk. This prior has 0.025, 0.5, 0.975 quantiles of 0.00098, 0.45, 5.0.

The addition of a non-integer also highlights that some adjustment has been made!

This prior is contributing information equivalent to observing an expected number of 0.5 and ‘half a case’.

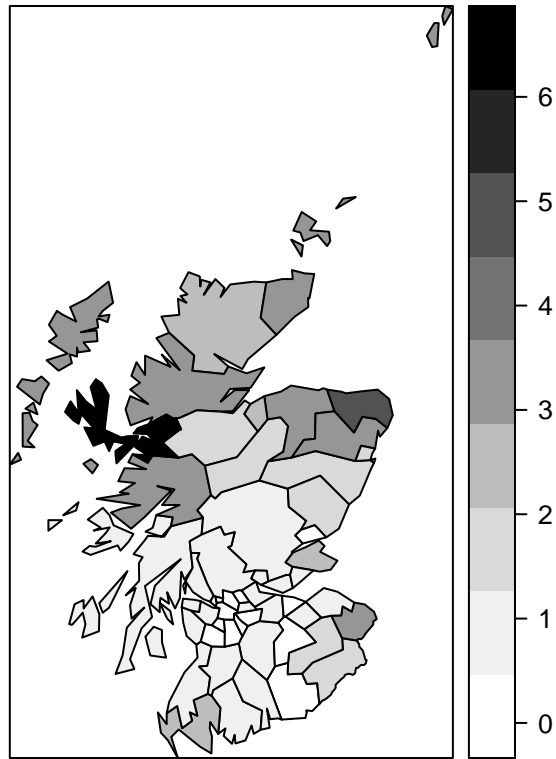
SMR estimates with adjustment. We create estimates adjusted for zeroes: if the number of cases is equal to zero both the number of cases and the expecteds are increased by 0.5.

```
Ystar <- ifelse(scotdata$cases==0,0.5,scotdata$cases)
Estar <- ifelse(scotdata$cases==0,scotdata$expected+0.5,scotdata$expected)
SMRstar <- Ystar/Estar
alphastar <- log(SMRstar)
varalphastar <- 1/(SMRstar*Estar)
SMRlower <- exp(alphastar-1.96*sqrt(varalphastar))
SMRupper <- exp(alphastar+1.96*sqrt(varalphastar))
SMRwidth <- SMRupper - SMRlower
scotdata$SMRstar <- SMRstar
scotdata$Estar <- Estar
scotdata$SMRlower <- SMRlower
```

```
scotdata$SMRupper <- SMRupper
scotdata$SMRwidth <- SMRwidth
smap <- SpatialPolygonsDataFrame(smap, scotdata, match.ID = TRUE)
```

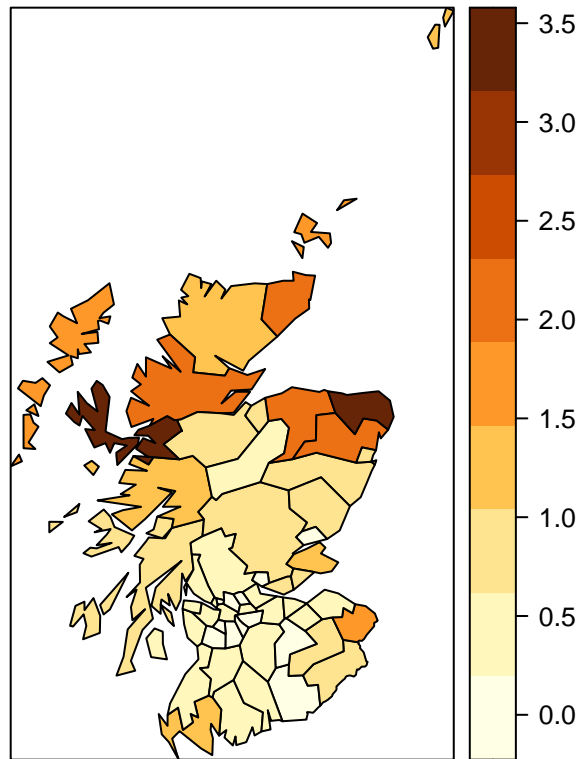
Point estimates with adjustment

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

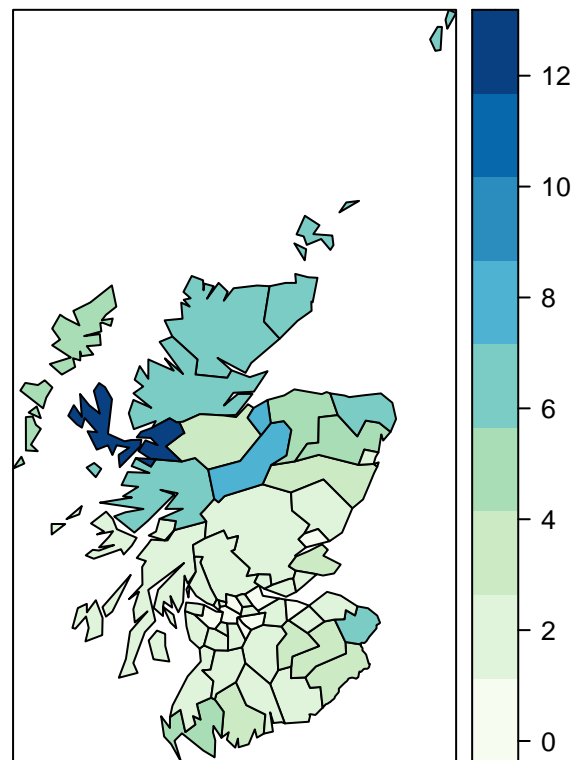


Lower and upper confidence intervals with adjustment.

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

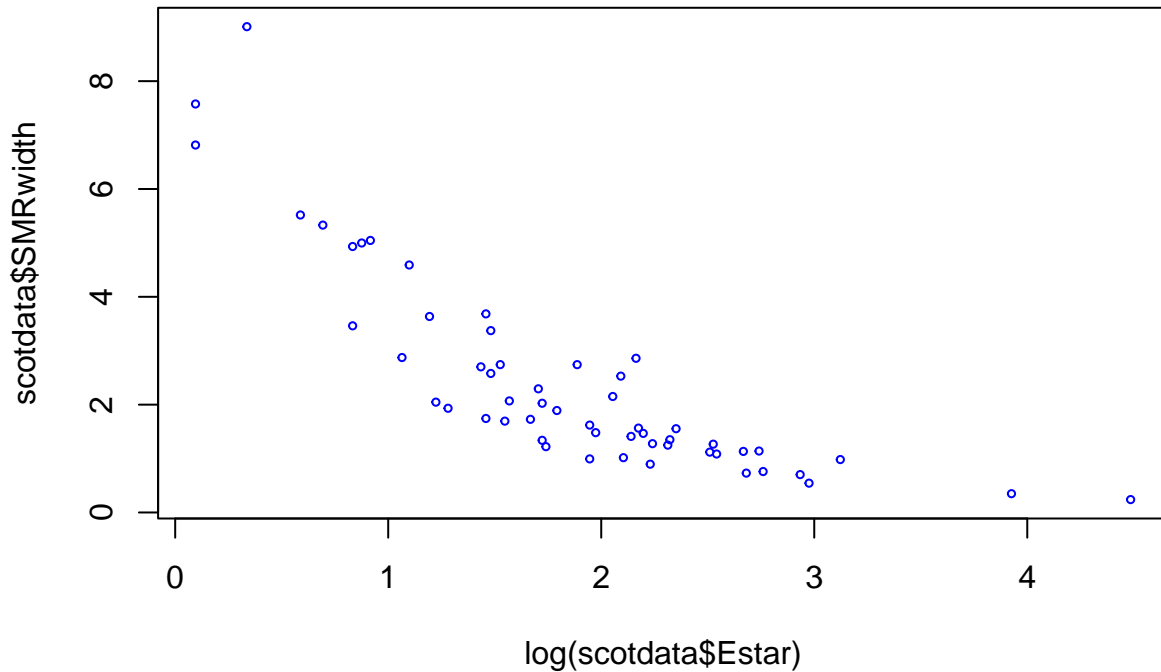


```
spplot(smap,zcol="SMRupper",col.regions=brewer.pal(9,"GnBu"),cuts=8)
```



The width decreases with increasing expected numbers!

```
plot(scotdata$SMRwidth~log(scotdata$Estar),col="blue",cex=.5)
```



## Poisson-Lognormal non-spatial smoothing 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:

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

where  $e_i$  are area-specific random effects that capture the residual or unexplained (log) relative risk of disease in area  $i$ ,  $i = 1, \dots, n = 56$ .

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,  $\tau_e$ ).

This model gives rise to the posterior distribution;

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

The full posterior is an  $(n + 2)$ -dimensional distribution and INLA by default produces summaries of the univariate posterior distributions for  $\beta_0$  and  $\tau_e$ .

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

## INLA for the Poisson-Lognormal model

We fit the Poisson-Lognormal model to the Scottish lip cancer data.

We first show a fit with no prior specifications given.

```
# Fit Poisson-lognormal model in INLA:
scotland.fit0 <- inla(Counts ~ 1 + f(Region, model="iid"),
  data=Scotland, family="poisson", E=E)
scotland.fit0$summary.fixed[,1:5]
##               mean          sd 0.025quant   0.5quant 0.975quant
```



```
## (Intercept) 0.08188339 0.1156166 -0.1498343 0.08333146 0.3054007
scotland.fit0$summary.hyper[,1:5]
##              mean              sd 0.025quant 0.5quant 0.975quant
## Precision for Region 1.840596 0.4680397 1.080752 1.787635 2.888909
```

A sanity check:

```
beta0est <- scotland.fit0$summary.fixed[,4]
sigma2est <- 1/scotland.fit0$summary.hyper[,4]
exp(beta0est+0.5*sigma2est)
## [1] 1.43768
```

Now we place a prior on the precision and ask for fitted values to be computed.

Note:

- The specification of the penalized complexity prior (Simpson et al 2017) for the precision  $\tau_e = \sigma_e^{-2}$ . Here we specify that there is a 5% chance that the standard deviation  $\sigma_e$  is greater than 1. The end of these notes contains a brief description of penalized complexity (PC) priors.

```
# Fit Poisson-lognormal model in INLA with prior specified
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 two lines give us calculated fitted values
control.predictor = list(compute = TRUE),
control.compute = list(return.marginals.predictor = TRUE))

scotland.fit1$summary.fixed[,1:5]
##              mean              sd 0.025quant 0.5quant 0.975quant
## (Intercept) 0.08067019 0.1166239 -0.1529431 0.08209424 0.3062218
scotland.fit1$summary.hyper[,1:5]
##              mean              sd 0.025quant 0.5quant 0.975quant
## Precision for Region 1.797849 0.4505067 1.064012 1.747348 2.805593
```

Very little sensitivity to the prior on the precision.

Let's look at the potential output:

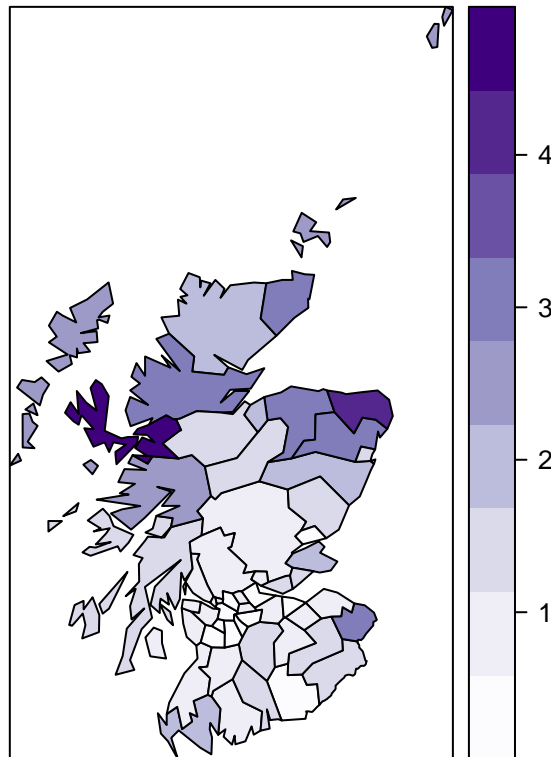
```
names(scotland.fit1)
## [1] "names.fixed"           "summary.fixed"
## [3] "marginals.fixed"       "summary.lincomb"
## [5] "marginals.lincomb"     "size.lincomb"
## [7] "summary.lincomb.derived" "marginals.lincomb.derived"
## [9] "size.lincomb.derived"  "mlik"
## [11] "cpo"                   "gcpo"
## [13] "po"                    "waic"
## [15] "residuals"            "model.random"
## [17] "summary.random"        "marginals.random"
## [19] "size.random"           "summary.linear.predictor"
## [21] "marginals.linear.predictor" "summary.fitted.values"
## [23] "marginals.fitted.values" "size.linear.predictor"
## [25] "summary.hyperpar"      "marginals.hyperpar"
## [27] "internal.summary.hyperpar" "internal.marginals.hyperpar"
## [29] "offset.linear.predictor" "model.spde2.blc"
## [31] "summary.spde2.blc"     "marginals.spde2.blc"
## [33] "size.spde2.blc"       "model.spde3.blc"
```

```
## [35] "summary.spde3.blc"      "marginals.spde3.blc"
## [37] "size.spde3.blc"        "logfile"
## [39] "misc"                   "dic"
## [41] "mode"                   "joint.hyper"
## [43] "nhyper"                 "version"
## [45] "Q"                      "graph"
## [47] "ok"                     "cpu.used"
## [49] "all.hyper"              ".args"
## [51] "call"                   "model.matrix"
```

We now extract the posterior medians of the log relative risks.

We now map the posterior medians of the relative risks.

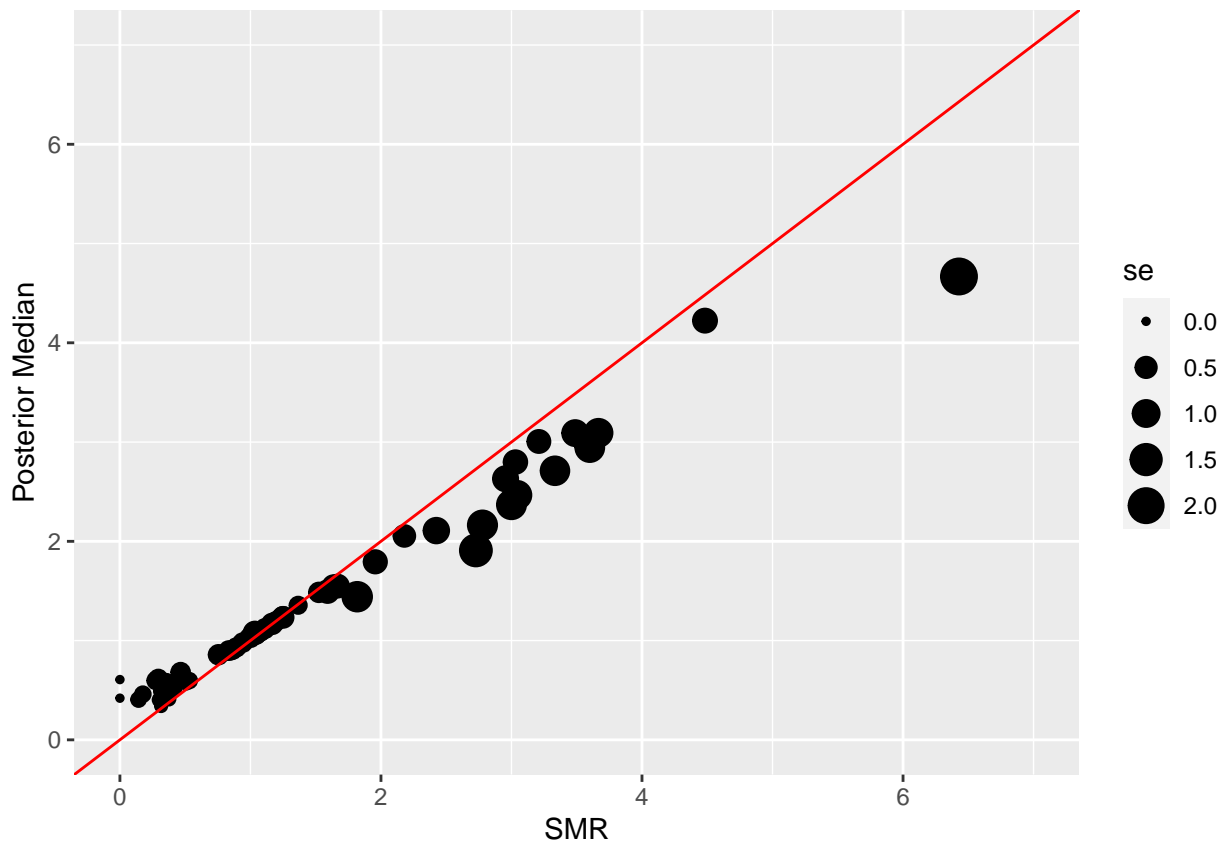
```
scotdata$fit1fitted <- scotland.fit1$summary.fitted.values$`0.5quant`
smap <- SpatialPolygonsDataFrame(smap,scotdata,match.ID = TRUE)
spplot(smap,zcol="fit1fitted",col.regions=brewer.pal(9,"Purples"),cuts=8)
```



Now compare the medians with the SMRs - we see some shrinkage, particularly for the low and high SMRs that have relatively large standard errors.

The standard errors of zero are artifacts of the SMRs being estimated as zero when  $Y_i = 0$ .

```
se <- sqrt(scotdata$SMR/scotdata$expected)
ggplot(data.frame(pmedian=scotland.fit1$summary.fitted.values$`0.5quant`,SMR=scotdata$SMR),
  aes(y=pmedian,x=SMR,size = se)) + geom_point() + labs(y="Posterior Median",x="SMR") + geom_abline
```

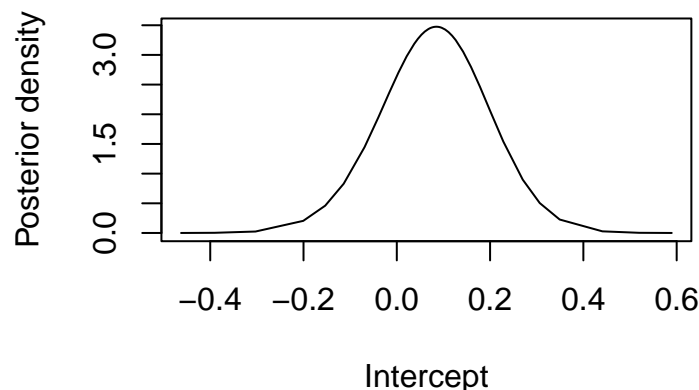


```
summary(scotland.fit1)
##
## Call:
##   c("inla.core(formula = formula, family = family, contrasts = contrasts,
##   ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##   scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##   ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##   verbose, ", " lincomb = lincomb, selection = selection, control.compute
##   = control.compute, ", " control.predictor = control.predictor,
##   control.family = control.family, ", " control.inla = control.inla,
##   control.fixed = control.fixed, ", " control.mode = control.mode,
##   control.expert = control.expert, ", " control.hazard = control.hazard,
##   control.lincomb = control.lincomb, ", " control.update =
##   control.update, control.lp.scale = control.lp.scale, ", "
##   control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##   ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##   num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##   working.directory = working.directory, ", " silent = silent, inla.mode
##   = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##   .parent.frame)")
## Time used:
##   Pre = 1.32, Running = 0.138, Post = 0.00605, Total = 1.47
## Fixed effects:
##           mean      sd 0.025quant 0.5quant 0.975quant  mode kld
## (Intercept) 0.081 0.117    -0.153   0.082     0.306 0.085  0
##
## Random effects:
```

```
## Name      Model
## Region IID model
##
## Model hyperparameters:
##           mean      sd 0.025quant 0.5quant 0.975quant mode
## Precision for Region 1.79 0.453      1.06      1.74      2.81 1.65
##
## Marginal log-Likelihood: -185.49
## is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
expbeta0med <- scotland.fit1$summary.fixed[4] # intercept
sdmed <- 1/sqrt(scotland.fit1$summary.hyperpar[4]) # sd
```

We examine the posterior marginal distribution for the intercept  $\beta_0$ .

```
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

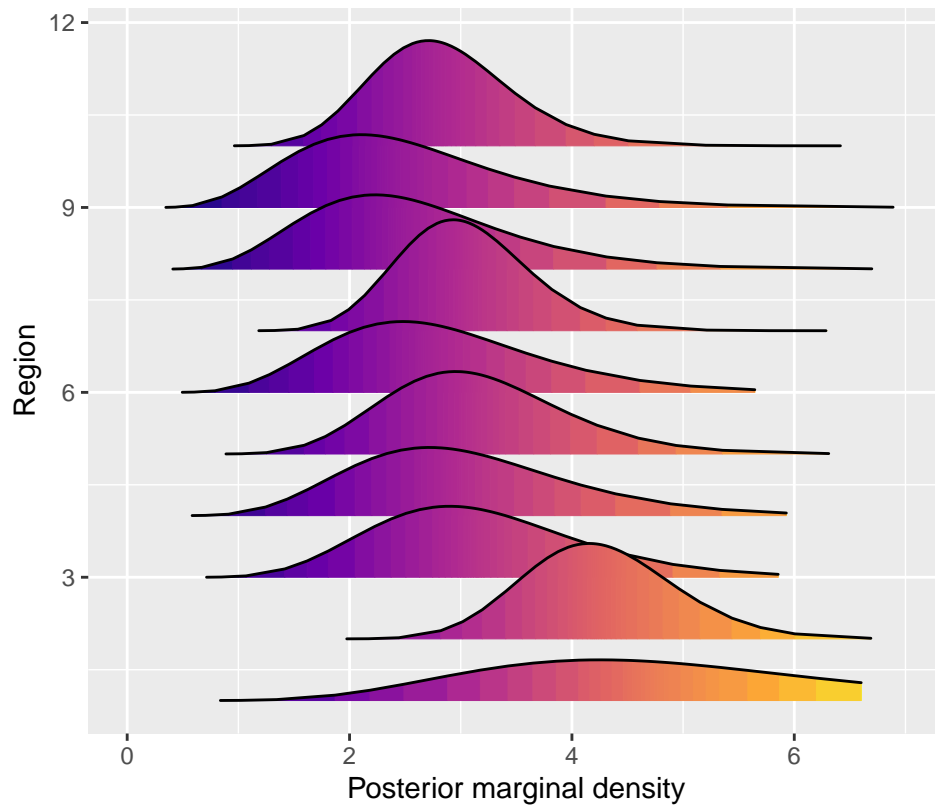
A function to extract a specified marginal for all regions from an INLA model

```
# 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)
}
```

We display ridgeplots for marginal posterior RRs in regions 1–10:

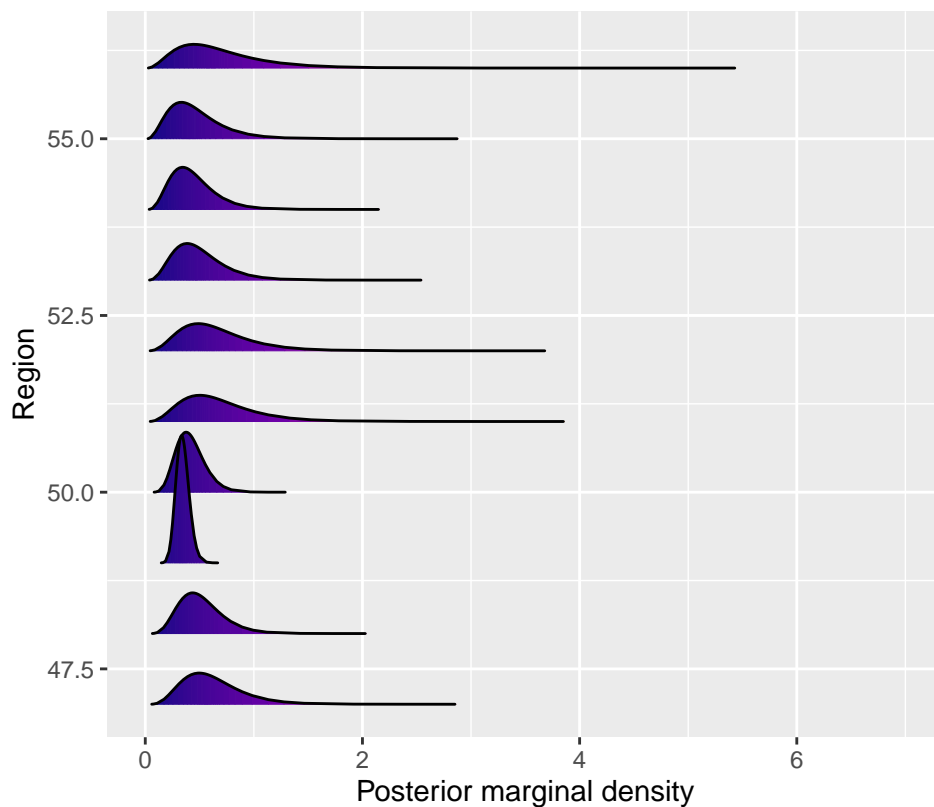
```
marginal_of_interest <- scotland.fit1$marginals.fitted.values
post_dens <- extract_marginals_to_plot(marginal_of_interest)
# we use the ggribes package to plot the marginals for first 28 Regions
ggplot(data = post_dens[post_dens$Region <= 10,],
  aes(x = x, y = Region, height = y, group = Region, 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')
```



Next, ridgeplots for marginal posterior RRs in regions 47–56

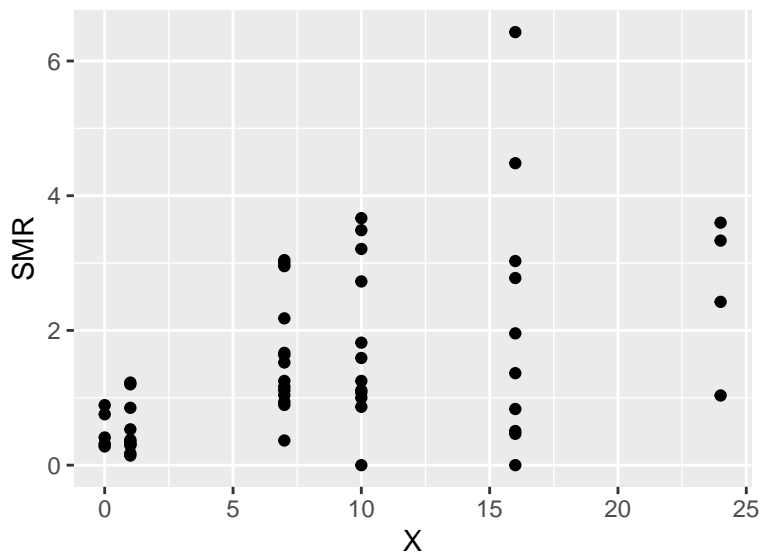
```
# we use the ggridges package to plot the marginals for last 10 Regions
ggplot(data = post_dens[post_dens$Region > 46,],
       aes(x = x, y = Region, height = y, group = Region, fill = ..x..)) +
  geom_density_ridges_gradient(stat = "identity", alpha = 0.5) +
  scale_fill_viridis_c(option = "C")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



### An excess-Poisson model with 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")
```



We fit a quasi-likelihood model with

$$\begin{aligned} E[Y_i] &= E_i \exp(\beta_0 + \beta_1 x_i) \\ \text{var}(Y_i) &= \kappa \times E[Y_i] \end{aligned}$$

This model allows for excess-Poisson variation (overdispersion) via  $\kappa$ , but does not allow for spatial dependence.

```

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$X
##  0.15418099  0.01320769
summary(modQL)
##
## Call:
## glm(formula = Scotland$Counts ~ Scotland$X, family = "quasipoisson",
##      offset = log(Scotland$E))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.7632  -1.2156   0.0967   1.3362   4.7130
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.54227    0.15418  -3.517 0.000893 ***
## Scotland$X   0.07373    0.01321   5.583 7.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.917964)
##
##      Null deviance: 380.73  on 55  degrees of freedom
## Residual deviance: 238.62  on 54  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5

```

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!).

The overdispersion is estimated as  $\hat{\kappa} = 4.9$ , which is considerable. Large excess-Poisson variation implies important missing covariates/confounders, and if these have spatial structure, then this will lead to strong spatial dependence (though we emphasize that the quasi-Poisson model does not account for this).

## Poisson-Lognormal non-spatial model with covariates

We now fit the three-stage model:

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

```
# 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$summary.fixed[1:5]
##              mean          sd 0.025quant  0.5quant 0.975quant
## (Intercept) -0.49195238 0.15962151 -0.8112866 -0.49017790 -0.1826872
## X           0.06836862 0.01425107  0.0403151  0.06836341  0.0964551
scotland.fit1X$summary.hyperpar[1:5]
##              mean          sd 0.025quant  0.5quant 0.975quant
## Precision for Region 2.937814 0.8405221  1.640496 2.824738  4.859334
```

## Poisson-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.

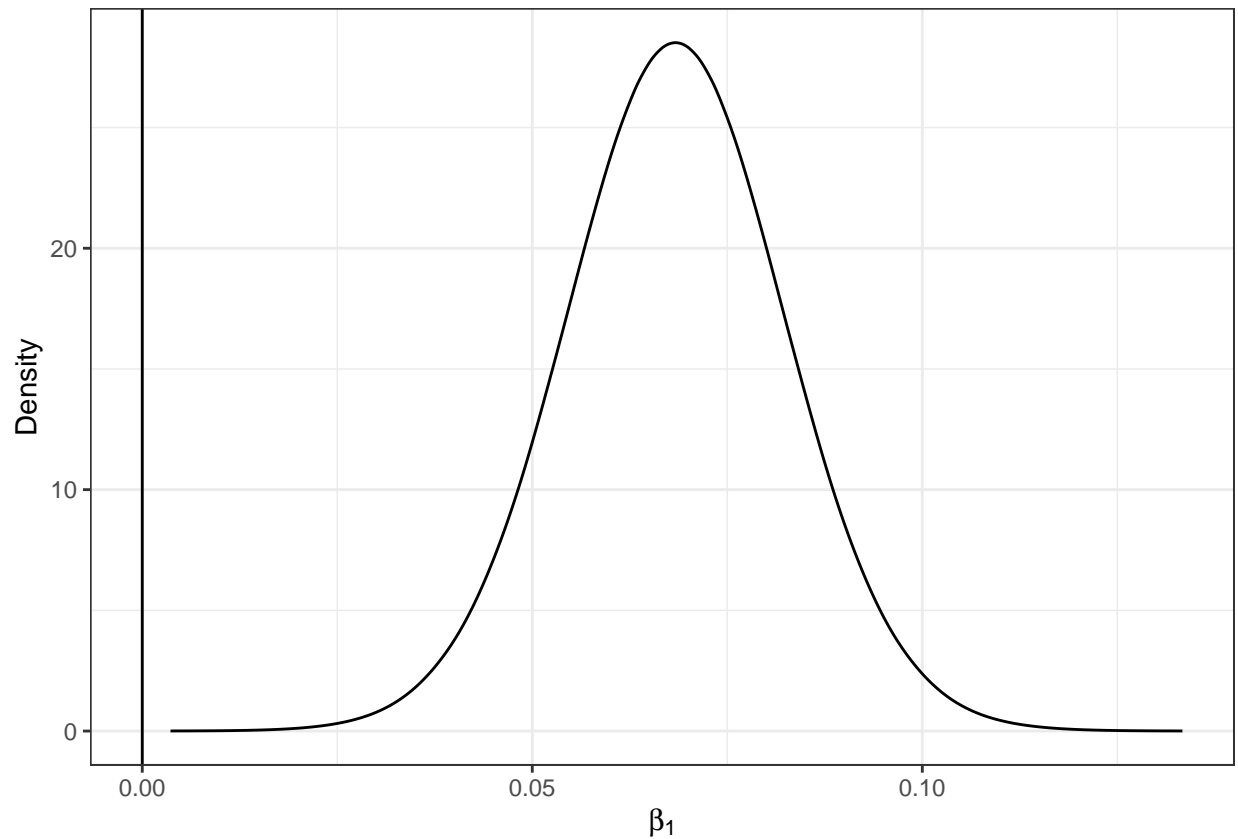
```
scotland.fit1X$summary.fixed[2,1:5]
##              mean          sd 0.025quant  0.5quant 0.975quant
## X 0.06836862 0.01425107  0.0403151  0.06836341  0.0964551
exp(scotland.fit1X$summary.fixed[2,3:5])
##      0.025quant 0.5quant 0.975quant
## X    1.041139 1.070754    1.10126
```

Note that we only exponentiate the quantiles of the posterior – the mean and variance cannot be legally exponentiated to give something useful.

Let's look at the posterior marginal for the log relative risk  $\beta_1$ .

```
marginal <- inla.ssmarginal(scotland.fit1X$marginals.fixed$X)
marginal <- data.frame(marginal)
ggplot(marginal, aes(x = x, y = y)) + geom_line() +
  labs(x = expression(beta[1]), y = "Density") +
  geom_vline(xintercept = 0, col = "black") + theme_bw()
```





## Parameter interpretation

```
scotland.fit1X$summary.fixed[1:5]
##              mean          sd 0.025quant    0.5quant 0.975quant
## (Intercept) -0.49195238 0.15962151 -0.8112866 -0.49017790 -0.1826872
## X           0.06836862 0.01425107  0.0403151  0.06836341  0.0964551
```

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  $\text{median}(\exp(\beta_1)|y) = \exp(0.068) = 1.07$ . This latter calculation exploits the fact that we can transform quantiles<sup>1</sup>

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.

```
scotland.fit1X$summary.fixed[1:5]
##              mean          sd 0.025quant    0.5quant 0.975quant
## (Intercept) -0.49195238 0.15962151 -0.8112866 -0.49017790 -0.1826872
## X           0.06836862 0.01425107  0.0403151  0.06836341  0.0964551
scotland.fit1X$summary.hyperpar[1:5]
##              mean          sd 0.025quant    0.5quant 0.975quant
## Precision for Region 2.937814 0.8405221  1.640496 2.824738  4.859334
```

The posterior median of  $\sigma_e$  is 0.594991699097777 and a 95% interval is

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

[0.453640313463064, 0.780750767895544]

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

A posterior median of a 95% RRR interval is

$$\begin{aligned} & [\exp(-1.96 \times \text{median}(\sigma_e)), \exp(1.96 \times \text{median}(\sigma_e))] \\ &= [\exp(-1.96 \times 0.595), \exp(1.96 \times 0.595)] = [0.31, 3.2] \end{aligned}$$

which is quite wide.

A more in depth analysis would examine the prior sensitivity to the prior on  $\tau_e$ .

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

## Poisson-Lognormal spatial model with a covariate

We now add spatial (ICAR) random effects to the model. We parameterize in terms of total variance and proportion that is spatial.

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

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

where  $x_i$  is the AFF in area  $i$ .

*Stage 2*: The random effects (prior distribution) is  $e_i | \sigma_e^2 \sim_{iid} N(0, \sigma_e^2)$  and the  $S_i$  are ICAR. The parameterization is in terms of the total variance  $\sigma_b^2$  and the proportion spatial  $\phi$ .

*Stage 3*: The hyperprior on the hyperparameters  $\beta_0, \beta_1, \sigma_b^2, \phi$ :

$$p(\beta_0, \beta_1, \sigma_b^2, \phi) = p(\beta_0)p(\beta_1)p(\sigma_b^2)p(\phi)$$

so that here we have assumed independent priors.

We need a graph file containing the neighbors.

```
# Spatial effects with covariate
download.file("http://faculty.washington.edu/jonno/SISMIDmaterial/scotland.graph", destfile = "scotland.graph")
```

Default specification:

```
formula <- Counts ~ 1 + X +
f(Region, model="bym2", graph="scotland.graph")
scotland.fit2default <- inla(formula, data=Scotland, family="poisson", E=E)
scotland.fit2default$summary.fixed[,1:5]
##           mean           sd 0.025quant 0.5quant 0.975quant
## (Intercept) -0.11658513 0.11055214 -0.335379373 -0.11601442 0.09893198
## X           0.02657229 0.01146762 0.003705277 0.02669152 0.04876789
scotland.fit2default$summary.hyper[,1:5]
##           mean           sd 0.025quant 0.5quant 0.975quant
## Precision for Region 4.8416204 1.46736313 2.5826544 4.633997 8.3072875
## Phi for Region       0.9329335 0.07714443 0.7153862 0.960283 0.9984225
```

We now place a penalized complexity prior on these two parameters and dot a few more i's.

```
formula <- Counts ~ 1 + X +
f(Region, model="bym2", graph="scotland.graph",
  scale.model=T,
```

```

constr=T,
rankdef = 1,
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(return.marginals.predictor=TRUE, config = TRUE))

scotland.fit2$summary.fixed[,1:5]
##              mean              sd 0.025quant  0.5quant 0.975quant
## (Intercept) -0.11674503 0.11151048 -0.337560637 -0.11613351 0.10055457
## X              0.02641061 0.01157946  0.003325369  0.02652868 0.04883155
scotland.fit2$summary.hyper[,1:5]
##              mean              sd 0.025quant  0.5quant 0.975quant
## Precision for Region 4.667003 1.42948466  2.4674456 4.4647379  8.0435856
## Phi for Region       0.936116 0.07518093  0.7234007 0.9632192  0.9987093

```

Slight differences on the hyperparameters (total precision and proportion spatial) but nothing substantive.

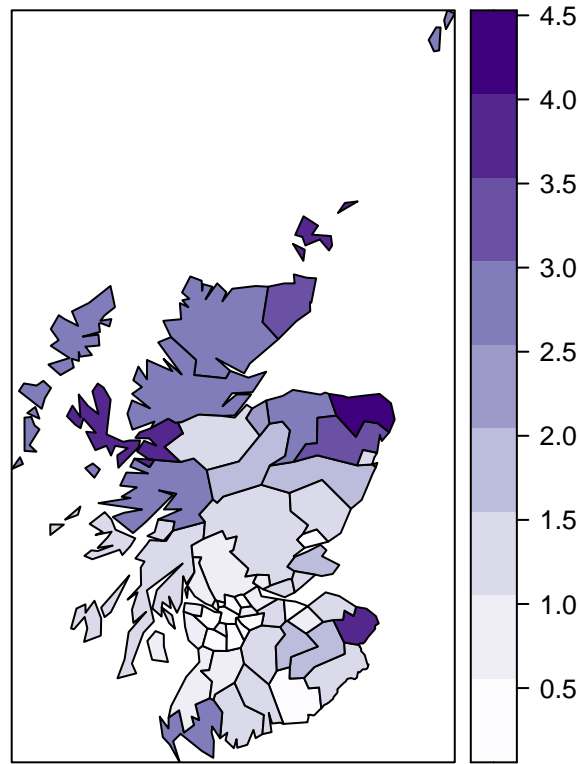
For the user-specified priors: The posterior median of the total standard deviation (on the log relative risk scale) is the posterior median of  $1/\sqrt{\tau_b}$  (where  $\tau_b$  is the precision), which is 0.4732624

The posterior median for the proportion of the residual variation that is spatial  $\phi$  is 0.9632192.

```

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

```



## Poisson-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.

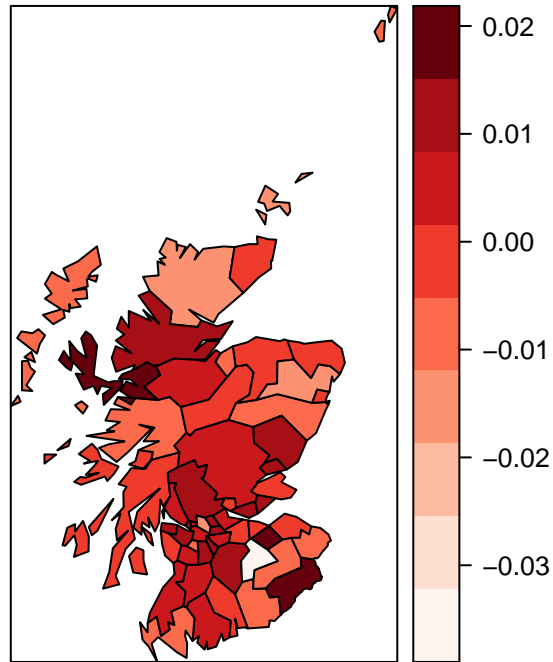
```
samp <- inla.posterior.sample(n = 1000, scotland.fit2)
samp_mat <- matrix(0, nrow = 1000, ncol = 2)
for (i in 1:1000) {
  samp_mat[i,] <- samp[[i]]$hyperpar[1:2]
}
scale_region <- mean(sqrt(samp_mat[,2])/sqrt(samp_mat[,1]))
```

## Poisson-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
scotdata$REsnonspat <- iid
scotdata$REsspat <- struct
```

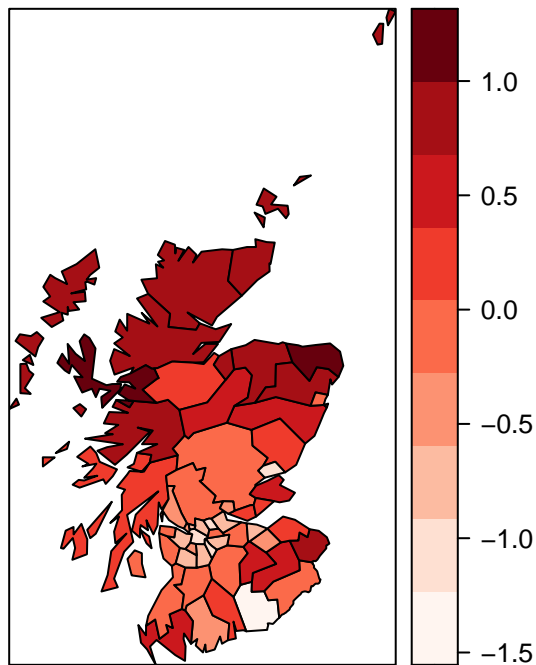
Non-spatial random effects:

```
smap <- SpatialPolygonsDataFrame(smap,scotdata,match.ID = TRUE)
spplot(smap,zcol="REsnonspat",col.regions=brewer.pal(9,"Reds"),cuts=8)
```



Spatial random effects:

```
smap <- SpatialPolygonsDataFrame(smap,scotdata,match.ID = TRUE)
spplot(smap,zcol="REsspat",col.regions=brewer.pal(9,"Reds"),cuts=8)
```



## Exceedance probabilities

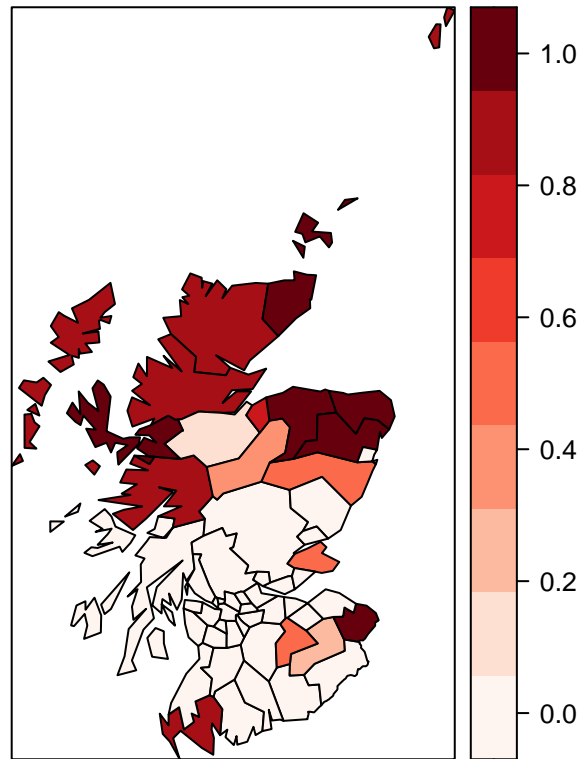
A useful summary is the posterior probability of exceedance of epidemiologically interesting thresholds.

Below we map the posterior probabilities

$$\Pr(\theta_i > 2|y),$$

for  $i = 1, \dots, 56$ .

```
exc <- sapply( scotland.fit2$marginals.fitted.values,  
FUN = function(marg){1 - inla.pmarginal(q = 2, marginal = marg)})  
smap$exc <- exc  
spplot(smap, zcol="exc", col.regions=brewer.pal(9, "Reds"), cuts=8)
```



## Spatial model: confounding by location

The command `plot(scotland.fit2)` provides plots of: marginal posterior distributions of  $\beta_0$ ,  $\beta_1$ ,  $\sigma_e^{-2}$ ,  $\sigma_S^{-2}$  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  $\beta_1$  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.

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

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  $\hat{S}_i$ .

A more complete analysis would address the sensitivity to the prior specifications on  $\sigma_e$  and  $\sigma_s$ .

## INLA Graph File Creation

The code below creates a neighborhood file for 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
download.file("http://faculty.washington.edu/jonno/SISMIDmaterial/wacounty.shp",destfile = "wacounty.shp")
download.file("http://faculty.washington.edu/jonno/SISMIDmaterial/wacounty.shx",destfile = "wacounty.shx")
download.file("http://faculty.washington.edu/jonno/SISMIDmaterial/wacounty.dbf",destfile = "wacounty.dbf")
countymap=readOGR(dsn=".",layer = "wacounty")
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/kpaulson/Repos/STAB", layer: "wacounty"
## with 39 features
## It has 6 fields
nb.map <- poly2nb(countymap)
nb2INLA("wacounty.graph",nb.map)
```

## Log relative risks as the outcome variable

Often the data arise in the form of observed rates or observed relative risks. For illustration, we imagine we had received the latter for Scotland, rather than the full data.

We model the log relative risk directly assuming they have a Gaussian distribution. We define  $Z_i = \log \hat{\theta}_i$  to emphasize that the observed data are now taken to be the log relative risks.

Recall that if any of the counts  $Y_i = 0$  (which would result in a relative risk of zero and a standard error of zero), we can use the approximations  $Y_i^* = Y_i + 0.5$  and  $E_i^* = E_i + 0.5$  to calculate  $\hat{\theta}_i^* = \frac{Y_i^*}{E_i^*}$ . In these cases,  $Z_i = \log \hat{\theta}_i^*$ . For simplicity, we assume this has been done.

In INLA, we can fit the model (with \*'s if necessary)

$$Z_i = \log \left( \frac{Y_i}{E_i} \right) \sim N(\mu_i, \sigma^2)$$

where  $\mu_i = E[Z_i]$ .

INLA estimates the precision for the Gaussian observations,  $1/\sigma^2$ . We evaluate the variance of the observed “data” by using the Poisson variance assumption —the mean equals the variance (in general the standard error of the rate can be estimated in a variety of ways, including the jackknife).

Therefore, the  $Z_i$  have “known” variances that we can approximate using the Delta method (as we did previously) as

$$\widehat{\text{var}}(Z_i) = [E_i \exp(Z_i)]^{-1} = \frac{1}{E_i \widehat{\theta}_i}$$

with \*'s if needed.

## Modeling the log relative risk as normal

We calculate the log relative risks as  $Z_i = \log \hat{\theta}_i^*$  and their variances for the Scotland lip cancer data.

```
Scotland$Ystar <- ifelse(Scotland$Counts == 0,
                        Scotland$Counts + 0.5,
                        Scotland$Counts)
Scotland$Estar <- ifelse(Scotland$Counts == 0,
                        Scotland$E + 0.5,
                        Scotland$E)
Scotland$thetastar <- Scotland$Ystar/Scotland$Estar
Scotland$Z <- log(Scotland$thetastar)
Scotland$varZ <- 1/(Scotland$Estar * Scotland$thetastar)
Scotland$precZ <- 1/Scotland$varZ
```

We can fit a normal model for the log relative risks (the likelihood) in INLA with fixed normal precisions (equivalent to known variance) using the following code.

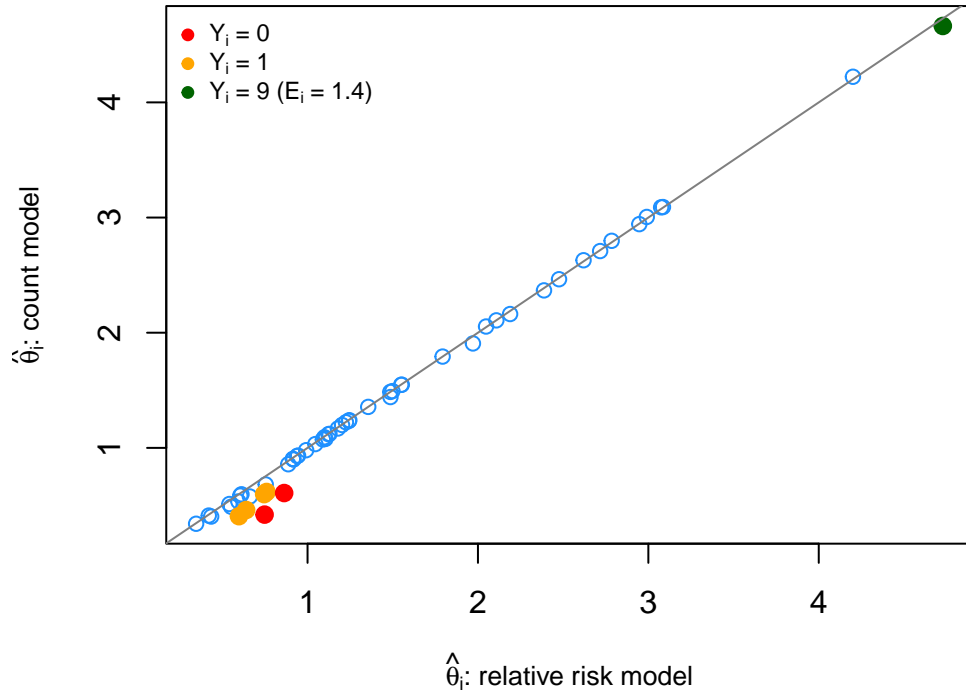
```
pcprec <- list(theta=list(prior='pc.prec',param=c(1,.05)))
scotland.fit3a <- inla(Z ~ 1 + f(Region, model="iid", hyper=pcprec),
  data = Scotland,family="gaussian",control.predictor=list(compute=TRUE),
  control.family = list(hyper = list(prec = list(initial = log(1), fixed=TRUE))),
  scale=precZ)
```

Note the scale=precZ which along with the previous line, fixes the measurement error variances.

We now compare the fits of the Poisson-Lognormal count outcome model and relative risk outcome models

```
par(mfrow = c(1,1), mar=c(5,4,1,1))
plot(scotland.fit1$summary.fitted.values[,4] ~ exp(scotland.fit3a$summary.fitted.values[,4]),
  col = "dodgerblue", ylab = expression(paste(hat(theta)[i],": count model")),
  xlab = expression(paste(hat(theta)[i],": relative risk model")), cex.lab = 0.85)
points(exp(scotland.fit3a$summary.fitted.values[,4][which(Scotland$Counts == 1)]),
  scotland.fit1$summary.fitted.values[,4][which(Scotland$Counts == 1)],
  col = "orange", pch = 19, cex = 1.15)
points(exp(scotland.fit3a$summary.fitted.values[,4][which(Scotland$Counts == 0)]),
  scotland.fit1$summary.fitted.values[,4][which(Scotland$Counts == 0)],
  col = "red", pch = 19, cex = 1.15)
points(exp(scotland.fit3a$summary.fitted.values[,4][which(Scotland$thetastar > 6)]),
  scotland.fit1$summary.fitted.values[,4][which(Scotland$thetastar > 6)],
  col = "darkgreen", pch = 19, cex = 1.15)
abline(0, 1, col = "gray50")
legend("topleft",legend = c(expression(paste(Y[i], " = ", 0)),
  expression(paste(Y[i], " = ", 1)),
  expression(paste(Y[i], " = ", 9, " (", E[i], " = ", 1.4, ")"))),
  pch = c(19, 19, 19), col = c("red", "orange", "darkgreen"), cex = 0.75, bty="n")
```





We fit a spatial normal log relative risk model with IID Normal random effects with a covariate  
Here, we add in a covariate and the spatial random effects

```
formula <- Z ~ 1 + I(X) +
  f(Region, model="bym2", graph="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.fit3 <- inla(formula, data=Scotland,
  family="gaussian", control.predictor=list(compute=TRUE),
  control.family = list(hyper = list(prec = list(initial = log(1), fixed=TRUE))),
  scale=precZ)
```

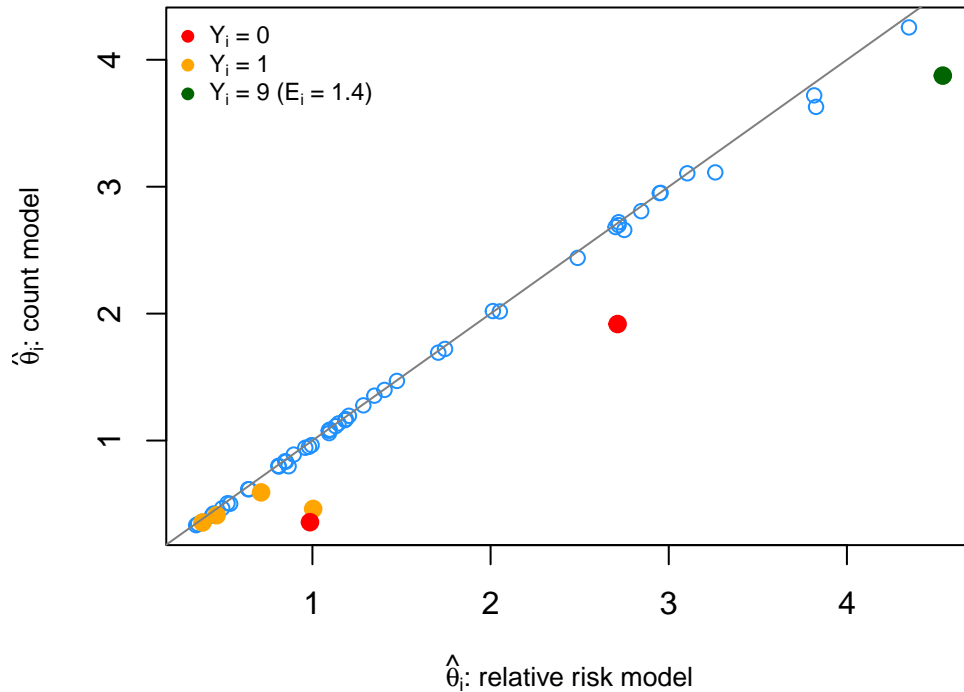
Comparison of spatial Poisson-Lognormal count outcome and relative risk outcome fits: differences in low and high extremes

```
par(mfrow = c(1,1), mar=c(5,4,1,1))
plot(scotland.fit2$summary.fitted.values[,4] ~ exp(scotland.fit3$summary.fitted.values[,4]),
  col = "dodgerblue", ylab = expression(paste(hat(theta)[i], ": count model")),
  xlab = expression(paste(hat(theta)[i], ": relative risk model")), cex.lab = 0.85)
points(exp(scotland.fit3$summary.fitted.values[,4][which(Scotland$Counts == 1)]),
  scotland.fit2$summary.fitted.values[,4][which(Scotland$Counts == 1)],
  col = "orange", pch = 19, cex = 1.15)
points(exp(scotland.fit3$summary.fitted.values[,4][which(Scotland$Counts == 0)]),
```

```

scotland.fit2$summary.fitted.values[,4][which(Scotland$Counts == 0)],
col = "red", pch = 19, cex = 1.15)
points(exp(scotland.fit3$summary.fitted.values[,4][which(Scotland$thetastar > 6)]),
scotland.fit2$summary.fitted.values[,4][which(Scotland$thetastar > 6)],
col = "darkgreen", pch = 19, cex = 1.15)
abline(0, 1, col = "gray50")
legend("topleft", legend = c(expression(paste(Y[i], " = ", 0)),
expression(paste(Y[i], " = ", 1)),
expression(paste(Y[i], " = ", 9, " (", E[i], " = ", 1.4, ")"))),
pch = c(19, 19, 19), col = c("red", "orange", "darkgreen"), cex = 0.75, bty="n")

```



Regression coefficient comparison: very similar estimates (and posterior uncertainty estimates) of the regression coefficients:

Count model:

```

scotland.fit2$summary.fixed
##           mean          sd  0.025quant    0.5quant 0.975quant
## (Intercept) -0.11674503 0.11151048 -0.337560637 -0.11613351 0.10055457
## X           0.02641061 0.01157946  0.003325369  0.02652868 0.04883155
##           mode          kld
## (Intercept) -0.11498077 7.696533e-07
## X           0.02676491 7.228436e-07

```

Relative risk model:

```

##           mean          sd  0.025quant    0.5quant 0.975quant
## (Intercept) -0.01461526 0.11083139 -0.232547542 -0.01452380 0.20275056
## I(X)         0.02495292 0.01139431  0.002371589  0.02502535 0.04712887
##           mode          kld
## (Intercept) -0.01439190 1.159633e-09
## I(X)         0.02517305 4.621065e-09

```

## Missing area data in Scotland

As an illustration we suppose that for the last area  $Y_{56}$  is unobserved – it is coded as NA (its value is zero in the data).

The missing value can be imputed with the spatial ICAR model helping in this respect.

If the count was missing because low (e.g., not released because less than 5) then this is informative and the following analysis is not appropriate.

```
Scotland$CountsNA <- Scotland$Counts
Scotland$CountsNA[56] <- NA
scotland.fitNA <- inla(CountsNA ~ 1 + I(X) +
f(Region, model="bym2",graph="scotland.graph",
  scale.model=T,
  constr=T,
  rankdef = 1,
  hyper=list(
    phi=list(
      prior="pc",
      param=c(0.5,0.5),
      initial=1),
    prec=list(
      prior="pc.prec",
      param=c(0.3/0.31,0.01),
      initial=5))) ,data=Scotland,
family="poisson",E=E,control.predictor=list(compute=TRUE,link=1))
```

```
summary(scotland.fitNA)
##
## Call:
## c("inla.core(formula = formula, family = family, contrasts = contrasts,
## ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
## scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
## ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
## verbose, ", " lincomb = lincomb, selection = selection, control.compute
## = control.compute, ", " control.predictor = control.predictor,
## control.family = control.family, ", " control.inla = control.inla,
## control.fixed = control.fixed, ", " control.mode = control.mode,
## control.expert = control.expert, ", " control.hazard = control.hazard,
## control.lincomb = control.lincomb, ", " control.update =
## control.update, control.lp.scale = control.lp.scale, ", "
## control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
## ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
## num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
## working.directory = working.directory, ", " silent = silent, inla.mode
## = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
## .parent.frame)")
## Time used:
## Pre = 8.55, Running = 0.213, Post = 0.0109, Total = 8.78
## Fixed effects:
##          mean      sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) -0.083 0.110      -0.300   -0.082      0.133 -0.082    0
## I(X)         0.024 0.011       0.002    0.024      0.046  0.025    0
##
## Random effects:
```

```
## Name      Model
## Region BYM2 model
##
## Model hyperparameters:
##           mean      sd 0.025quant 0.5quant 0.975quant  mode
## Precision for Region 4.992 1.474      2.704    4.790      8.457 4.410
## Phi for Region      0.948 0.064      0.766    0.971      0.999 0.999
##
## Marginal log-Likelihood: -126.76
## is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

From the graph file we see that area 56 has areas 2,3,4,5 as neighbors – we look at these values and see the SMRs are high, which explains why the predictive mean is high.

We include the prediction of the rate from the model in which the data are observed and see it is much lower.

We look at the last line of the graph file (for area 56):

```
56 4 2 3 4 5
```

We examine the SMRs from the 4 neighboring areas, and they are high.

The covariate value for area 56 (which is also used in the prediction) is mid-range.

We obtain the expected count of the missing area response and compare with the true response (which of course we know in this exercise)

```
scotland.fitNA$summary.fitted.values[56,]
##           mean      sd 0.025quant 0.5quant 0.975quant  mode
## fitted.Predictor.56 3.271555 1.257975  1.491034 3.051826  6.336698 2.689721
Scotland$E[56]
## [1] 1.8
Scotland$X[56]
## [1] 10
set56 <- c(2,3,4,5)
Scotland$Counts[set56]
## [1] 39 11  9 15
Scotland$E[set56]
## [1] 8.7 3.0 2.5 4.3
Scotland$Counts[set56]/Scotland$E[set56]
## [1] 4.482759 3.666667 3.600000 3.488372
# Compare with fit from model in which Y[56]=0 is used as observed
scotland.fit2$summary.fitted.values[56,]
##           mean      sd 0.025quant 0.5quant 0.975quant  mode
## fitted.Predictor.56 1.988273 0.6236627 0.9762036 1.917764  3.408063 1.785137
```

Much lower predictor because the zero pulls down.

In the model

$$Y_i|\theta_i \sim \text{Poisson}(E_i\theta_i),$$

the fitted values/predictions are for  $\theta_i$ , so no expected number and no Poisson sampling (so we're predicting the the relative risk).

We confirm that the quantiles of the fitted values are the exponentiated predictions.

```
scotland.fitNA$summary.fitted.values[56,c(3:5)]
##           0.025quant 0.5quant 0.975quant
```

```
## fitted.Predictor.56 1.491034 3.051826 6.336698
exp(scotland.fitNA$summary.linear.predictor[56,c(3:5)])
## 0.025quant 0.5quant 0.975quant
## Predictor.56 1.491034 3.051826 6.336698
```

## Censored Poisson Data in Scotland

Sometimes, data come in a form where the smallest counts are censored and we only know that a count is below a certain threshold. For example, due to privacy concerns data with small counts in.

To illustrate, we will censor the counts in the Scotland data that are below 2.

```
Scotland$CountsCen <- ifelse(Scotland$Counts < 2, 0, Scotland$Counts)
```

We now show how to implement the censored Poisson model in INLA.

We will fit a Poisson-Lognormal model with a covariate and spatial REs, using PC priors.

```
# fit a Lognormal model with a censored Poisson family
# with the censoring interval set to anything less than 2
scotland.fit.cen <- inla(CountsCen ~ 1 + I(X) +
  f(Region, model="bym2", graph="scotland.graph",
    scale.model=T,
    constr=T,
    rankdef = 1,
    hyper=list(
      phi=list(
        prior="pc",
        param=c(0.5,0.5),
        initial=1),
      prec=list(
        prior="pc.prec",
        param=c(0.3/0.31,0.01),
        initial=5))),
  family = "cenpoisson",
  control.family = list(cenpoisson.I = c(0,1)),
  E=E,
  data = Scotland,
  control.predictor=list(compute=TRUE, link=1))
```

We examine the regression coefficients for the censored and uncensored data.

Censored data:

```
scotland.fit.cen$summary.fixed
## mean sd 0.025quant 0.5quant 0.975quant
## (Intercept) -0.12119696 0.11044931 -0.340026585 -0.12054694 0.09389475
## I(X) 0.02723674 0.01141538 0.004484709 0.02735192 0.04934206
## mode kld
## (Intercept) -0.11932214 7.261633e-07
## I(X) 0.02758365 7.253799e-07
```

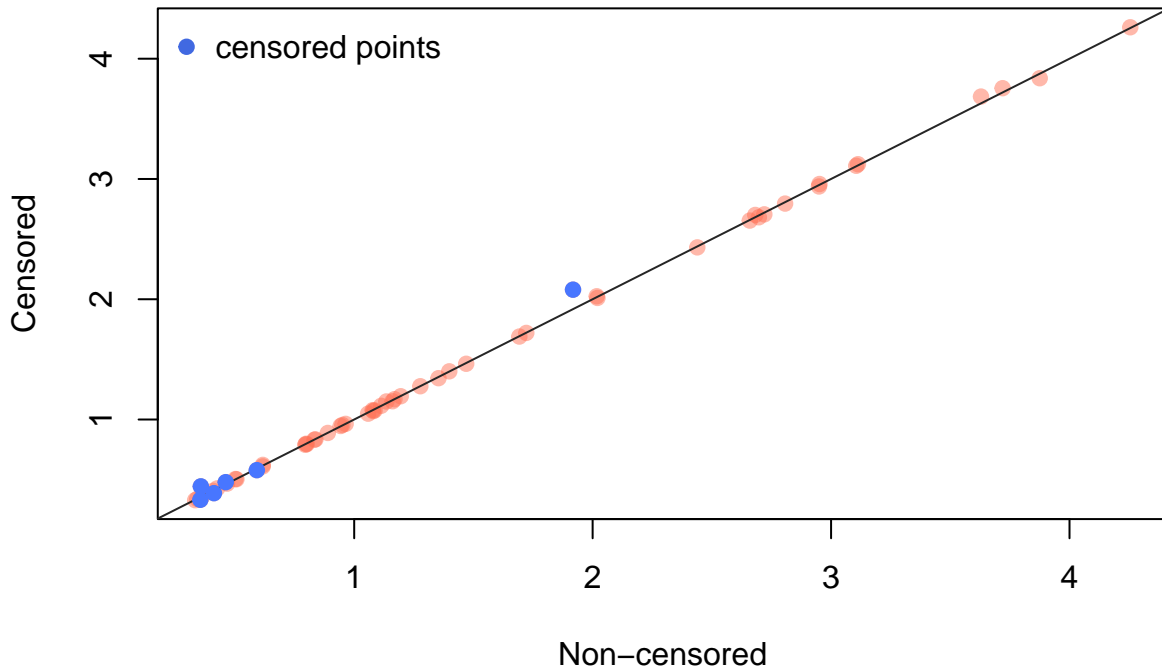
Uncensored data:

```
scotland.fit2$summary.fixed
## mean sd 0.025quant 0.5quant 0.975quant
## (Intercept) -0.11674503 0.11151048 -0.337560637 -0.11613351 0.10055457
```

```
## X          0.02641061 0.01157946 0.003325369 0.02652868 0.04883155
##          mode          kld
## (Intercept) -0.11498077 7.696533e-07
## X          0.02676491 7.228436e-07
```

We compare fitted values with the model fitted to the uncensored data

```
cen_est <- scotland.fit.cen$summary.fitted.values$`0.5quant`
non_cen_est <- scotland.fit2$summary.fitted.values$`0.5quant`
par(mfrow=c(1,1))
plot(cen_est ~ non_cen_est,
     col = alpha("coral1",0.5),pch=19,
     xlab="Non-censored",ylab="Censored")
abline(0,1,col="gray15")
points(non_cen_est[Scotland$CountsCen==0],
       cen_est[Scotland$CountsCen==0],
       col = "royalblue1",pch=19)
legend("topleft",c("censored points"),pch=c(19), col=c("royalblue"),bty="n")
```



## 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  $\lambda$ , 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  $\tau$ , 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).

The PC prior for  $\phi$  in the BYM2 random effect is more complex, and does not have a nice distribution. However, it retains the interpretation of a PC prior, which is why it is useful. See details in Simpson et al. (2017).