Bayesian Regression of Radon Levels

Greg Johnson

```
require(geoR)
require(mvtnorm)
```

A city collects radon levels from houses in three counties, with some measurements coming from the basement and some not.

```
dat = data.frame(BE = c(rep(1, 14), rep(0, 27)), C = c(rep(0,
    14), rep(1, 14), rep(0, 13)), G = c(rep(0, 28), rep(1, 13)),
    basement = c(1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0,
        1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1,
        1, 1, 1, 1, 1, 1, 1, 1), logradon = log(c(5, 13, 7.2,
        6.8, 12.8, 5.8, 9.5, 6, 3.8, 14.3, 1.8, 6.9, 4.7, 9.5,
        0.9, 12.9, 2.6, 3.5, 26.6, 1.5, 13, 8.8, 19.5, 2.5, 9,
        13.1, 3.6, 6.9, 14.3, 6.9, 7.6, 9.8, 2.6, 43.5, 4.9,
        3.5, 4.8, 5.6, 3.5, 3.9, 6.7)))

Y = dat[, 5]
X = as.matrix(dat[, 1:4])
```

We want to predict radon levels based on county and whether the measurement was taken in basement. We don't necessarily have prior information about the predictive quality of our predictors but we would like to conduct a Bayesian regression nonetheless, mainly for our Bayesian model inferences e.g. credible intervals.

Our basic OLS model assumes that the errors are independent, zero-mean, constant-variance, and Gaussian and that the expectation of our outcome, logarithm of radon, is linear in our predictors:

$$\mathbf{y}|\boldsymbol{\beta},\sigma^2,X \sim \mathrm{N}(X\boldsymbol{\beta},\sigma^2 I_n)$$

We assign a noninformative prior distribution that is uniform on $(\boldsymbol{\beta},\sigma^2)$:

$$p(\boldsymbol{\beta},\sigma^2|X) \propto \sigma^{-2}$$

For the joint posterior we first derive the conditional posterior of $\boldsymbol{\beta}$ given σ^2, then we derive the marginal posterior of σ^2.

$$\begin{gathered}
\boldsymbol{\beta}|\sigma,\mathbf{y} \sim \mathrm{N}(\hat{\boldsymbol{\beta}},\sigma^2 V_\beta) \\
\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} \\
V_\beta = (X^T X)^{-1} \\
\sigma^2|\mathbf{y} \sim \mathrm{Inv.}\chi^2(n-k,s^2) \\
s^2 = \frac{1}{n-k}(\mathbf{y}-X\hat{\boldsymbol{\beta}})^T(\mathbf{y}-X\hat{\boldsymbol{\beta}})
\end{gathered}$$

We can use a simple simulation strategy to sample from the posterior for $(\boldsymbol{\beta},\sigma^2)$. For $t = 1,...,T$

1. Sample σ^2 from its marginal posterior.
2. Sample $\boldsymbol{\beta}$ from its conditional posterior.

```
BayesRegression = function(Y, X, ndraws) {
    n = nrow(X)
    k = ncol(X)
    Betahat = solve(t(X) %*% X) %*% t(X) %*% Y
    Vbeta = solve(t(X) %*% X)

    post_draws = matrix(NA, ndraws, 5, dimnames = list(NULL,
        c("sig_sq", "beta_BE", "beta_C", "beta_G", "beta_base")))
    for (j in 1:ndraws) {
        sig_sq = as.numeric(rinvchisq(1, df = n - k, scale = 1/(n -
            k) * t(Y - X %*% Betahat) %*% (Y - X %*% Betahat)))
        beta = rmvnorm(1, mean = Betahat, sigma = sig_sq * Vbeta)
        post_draws[j, 1] = sig_sq
        post_draws[j, 2:5] = beta
    }
    return(post_draws)
}

post_draws = BayesRegression(Y, X, 1000)
```

For our posterior inference we will look at 95% credible intervals and MAP (maximum a posteriori) estimates.

```
CredInt95 = matrix(NA, 5, 3, dimnames = list(c("sig_sq", "beta_BE",
    "beta_C", "beta_G", "beta_base"), c("LB", "UB", "MAP")))

for (j in 1:ncol(post_draws)) {
    CredInt95[j, 1:2] = quantile(post_draws[, j], c(0.025, 0.975))
    d = density(post_draws[, j])
    i = which.max(d$y)
    CredInt95[j, 3] = d$x[i]
}

CredInt95
```

```
##                   LB        UB       MAP
## sig_sq     0.4279948 1.0031764 0.5683314
## beta_BE    0.8799319 2.3419624 1.5750288
## beta_C     0.9379483 2.1717298 1.4855279
## beta_G     0.8716718 2.2707872 1.5791374
## beta_base -0.2796897 0.9752356 0.3034346
```

Let's look at the posterior inference for our main effects individually.

1. There is a 95% chance that Blue Earth houses have radon levels between 2.5 and 11 (4.9 is the most likely) in the first floor and between 3.2 and 13.6 (6.1 is the most likely) in the basement In other words we are fairly certain that there is some amount of radon in Blue Earth houses whether in the basement or the first floor.

2. There is a 95% chance that Clay County houses have radon levels between 2.5 and 8.7 (5.0 is the most likely) in the first floor and between 3.2 and 9.4 (6.3 is the most likely) in the basement.

3. There is a 95% chance that Goodhue County houses have radon levels between 2.4 and 10.3 (5.2 is the most likely) in the first floor and between 3.1 and 12.7 (6.5 is the most likely) in the basement.

Suppose another house is sampled at random from Blue Earth. Let's consider radon levels in its first floor and basement separately.

1. Blue Earth, basement $\tilde{\boldsymbol{x}}_1 = (1, 0, 0, 1$

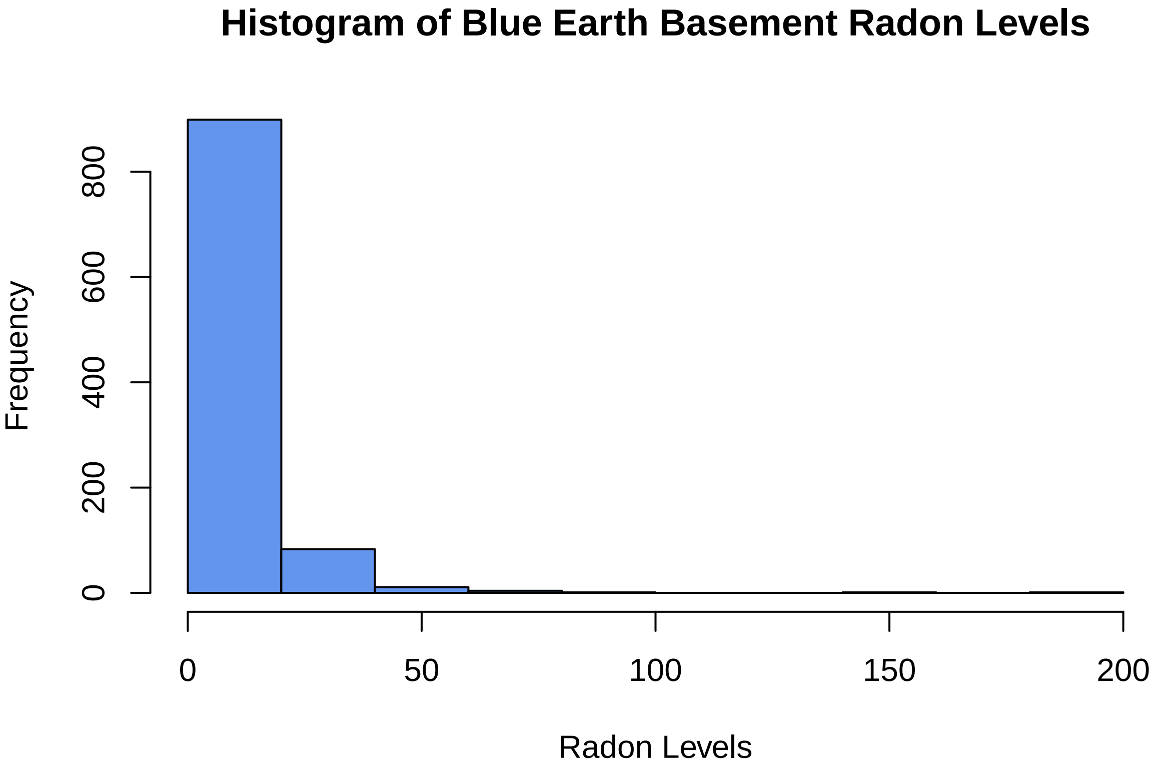
2. Blue Earth, first floor $\tilde{\boldsymbol{x}}_2 = (1, 0, 0, 0)$

```
postpred_BEbase = numeric(nrow(post_draws))
postpred_BEfirst = numeric(nrow(post_draws))
x1 = c(1, 0, 0, 1)
x2 = c(1, 0, 0, 0)

for (j in 1:nrow(post_draws)) {
    beta = post_draws[j, 2:5]
    sig_sq = post_draws[j, 1]
    postpred_BEbase[j] = rnorm(1, mean = t(x1) %*% beta, sd = sqrt(sig_sq))
    postpred_BEfirst[j] = rnorm(1, mean = t(x2) %*% beta, sd = sqrt(sig_sq))
}
```

Let's look at our simulation of a Blue Earth basement:

```
hist(exp(postpred_BEbase), main = "Histogram of Blue Earth Basement Radon Levels",
    xlab = "Radon Levels", col = "cornflowerblue")
```

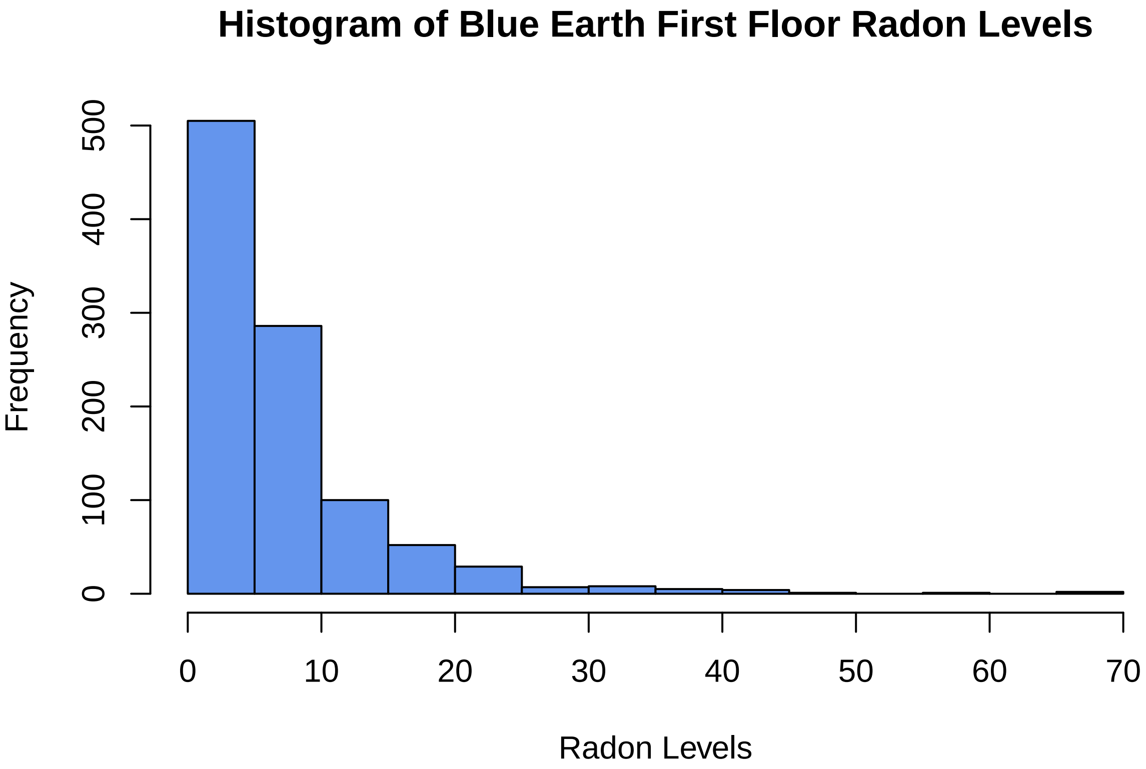


```
# 95% credible intervals
exp(quantile(postpred_BEbase, c(0.025, 0.975)))
```

```
##      2.5%     97.5%
##  1.466643 36.799682
```

Next we look at the first floor:

```
hist(exp(postpred_BEfirst), main = "Histogram of Blue Earth First Floor Radon Levels",
    xlab = "Radon Levels", col = "cornflowerblue")
```



```r
# 95% credible intervals
exp(quantile(postpred_BEfirst, c(0.025, 0.975)))
```

```
##     2.5%    97.5%
##  0.89669 27.31955
```

Our Bayesian regression tells us that for a Blue Earth home, there is a 95% chance that radon levels are between 1.3 and 45.6 in the basement, and between 1.0 and 31.8 in first floor.