Random effects ANOVA
Created by: AJ Richards
Last updated: November 28, 2014

Contents

1 Problem and data **3**

1.1 About . . . 3

1.2 Fixed and random effects ANOVA . . . 4

2 Fixed effects ANOVA **5**

2.1 Maximum likelihood analysis . . . 5

2.2 MCMC analysis . . . 5

3 Random effects ANOVA **8**

3.1 Maximum likelihood analysis . . . 8

3.2 MCMC analysis . . . 9

4 Reproducibility **11**

1 Problem and data

1.1 About

This is an example from Marc Kery's book "Introduction to WinBUGS for ecologists [1]. Here we explore fixed and random effects models with ANOVA (a t-test applied more than two groups). First we generate the data— with a number of populations of snakes (`ngroups`) each with an equal number of `nsample` and a single measured co-variate `svl`. From within the function generateData.R the following snippets show the parameters used to create the fixed-effects and random effects data sets. In this problem we are interested in whether or not the five population of snakes differ.

```
### generate data for fixed effects ###
ngroups <- 5
nsample <- 10
popMeans <- c(50,40,45,55,60)
sigma <- 3

n <- ngroups * nsample
resid <- rnorm(n,0,sigma)


### generate data for random effects ###
npop <- 10
nsample <- 12
n <- npop * nsample
popGrandMean <- 50
popSd <- 5
popMeans <- rnorm(n=npop,mean=popGrandMean,sd=popSd)
sigma <- 3
```

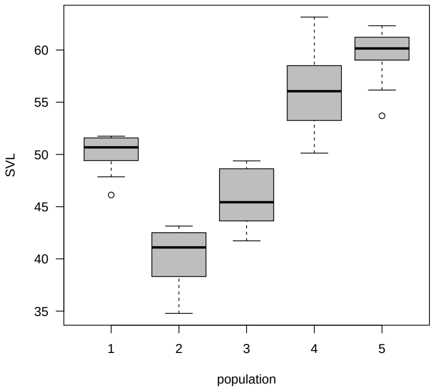

Figure 1.1: The distributions of the five populations of snakes with respect to snout-vent-length

1.2 Fixed and random effects ANOVA

In a random effects model the **effects** (or group means) are constrained to come from some distribution (usually Gaussian or Bernoulli). The means parameterization for the one-way ANOVA:

$$y_i = \alpha_{j(i)} + \epsilon_i \tag{1.1}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1.2}$$

$$\alpha_{j(i)} \sim \mathcal{N}(\mu, \tau^2) \tag{1.3}$$

y_i refers to the `svl` of snake i in population j. $\alpha_{j(i)}$ is the expected `svl` of a snake in population j and the residual ϵ_i is the random deviation of snake i from its population mean $\alpha_{j(i)}$. Eqn 1.3 is the key assumption that moves this from a fixed-effects to a random effects ANOVA. If we do not specify that the population means $\alpha_{j(i)}$ come from a distribution we are using a fixed-effects ANOVA. It is not always clear whether we should be using a fixed-effect or random-effects ANOVA, in fact statisticians have differing opinions. Random effects are often used for things like `year`, `month`, or `location`. Some opinions say the decision has to do with whether we want to generalize our conclusions to the larger (unsampled population. Another train of thought is that if we have reason to believe that there might be a difference between populations, but the populations themselves are quite similar then it may be appropriate to use random-effects. Another way to analyze these data is pool the samples. If we do this we assume that there are no effects. If we use fixed effects we assume the effects are completely independent and we can see random effects setup as a compromise between the two. In this example we will use Akaike information criterion (AIC) to help make a determination.

2 Fixed effects ANOVA

2.1 Maximum likelihood analysis

```r
data = read.csv("fe-svl.csv")
print("AVOVA-results")
fit = lm(data$y~as.factor(data$x))
print(anova(fit))
cat("\n")
print("Linear-model-effects-parameterization")
print(summary(fit)$coeff,dig=3)
cat("Sigma:", summary(fit)$sigma, "\n")
print(paste(’AIC’, AIC(fit)))
cat("\n")
print("Linear-model-means-parameterization")
fit = lm(data$y~as.factor(data$x)-1)
print(summary(fit)$coeff,dig=3)
cat("Sigma:", summary(fit)$sigma, "\n")
print(paste(’AIC’, AIC(fit)))
```

```
[1] "..."
[1] "AVOVA-results"
Analysis of Variance Table

Response: data$y
                Df Sum Sq Mean Sq F value Pr(>F)
as.factor(data$x) 4 2422.57 605.64 68.109 < 2.2e-16 ***
Residuals 45 400.15 8.89
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

[1] "Linear-model-effects-parameterization"
                 Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.11 0.943 53.14 2.99e-42
as.factor(data$x)2 -9.87 1.334 -7.40 2.57e-09
as.factor(data$x)3 -4.36 1.334 -3.27 2.05e-03
as.factor(data$x)4 6.10 1.334 4.57 3.77e-05
as.factor(data$x)5 9.43 1.334 7.07 8.06e-09
Sigma: 2.981976
[1] "AIC 257.884446651633"

[1] "Linear-model-means-parameterization"
                 Estimate Std. Error t value Pr(>|t|)
as.factor(data$x)1 50.1 0.943 53.1 2.99e-42
as.factor(data$x)2 40.2 0.943 42.7 4.81e-38
as.factor(data$x)3 45.7 0.943 48.5 1.69e-40
as.factor(data$x)4 56.2 0.943 59.6 1.83e-44
as.factor(data$x)5 59.5 0.943 63.1 1.41e-45
Sigma: 2.981976
[1] "AIC 257.884446651633"
```

Be reminded that R fits an effects parameterization of ANOVA by default so the effects are relative to the first population. This means that first population in `x` is set to zero and the relative differences are reported. The means parameterization is straight from the data. If you look above in the simulated data $\sigma = 3.0$ and the population means were 50,40,45,55 and 60. The null hypothesis is that the means are the same so the significant p-values reject that null.

2.2 MCMC analysis

With BUGS we generally fit a mean parameterization and calculate the differences between populations (effects) using derived quantities. See the script fixedEffectAnova.R for more details. Here is the output.

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 122

Initializing model

$BUGSoutput
Inference for Bugs model at "fe-anova.txt", fit using jags,
 3 chains, each with 5000 iterations (first 200 discarded), n.thin = 2
 n.sims = 7200 iterations saved
          mean  sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
alpha[1]  50.1 1.0  48.1  49.4  50.1  50.7  52.0    1  7200
alpha[2]  40.2 1.0  38.3  39.5  40.2  40.8  42.1    1  7200
alpha[3]  45.7 1.0  43.8  45.0  45.7  46.4  47.6    1  5300
alpha[4]  56.2 1.0  54.2  55.5  56.2  56.8  58.0    1  2700
alpha[5]  59.5 1.0  57.5  58.8  59.5  60.1  61.4    1  7200
deviance 252.5 3.8 247.3 249.7 251.8 254.6 261.8    1  2100
effe2     -9.9 1.4 -12.6 -10.8  -9.9  -9.0  -7.2    1  7200
effe3     -4.3 1.4  -7.1  -5.3  -4.3  -3.4  -1.6    1  7200
effe4      6.1 1.4   3.4   5.2   6.1   7.0   8.8    1  7200
effe5      9.4 1.4   6.7   8.5   9.5  10.4  12.2    1  7200
sigma      3.1 0.3   2.5   2.8   3.0   3.3   3.8    1  7200
test1    -29.7 2.0 -33.6 -31.1 -29.7 -28.4 -25.8    1  2600
test2     -2.8 2.4  -7.5  -4.4  -2.8  -1.2   2.0    1  4400

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 7.3 and DIC = 259.8
DIC is an estimate of expected predictive error (lower deviance is better).
```

We see that the results are essentially the same. Here are common plots used to examine the posterior.

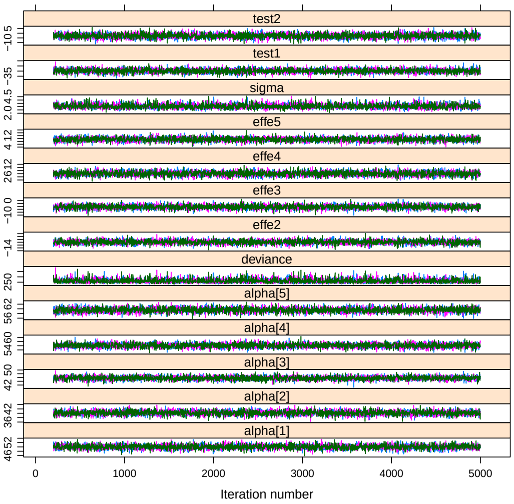

Figure 2.1: The MCMC chains

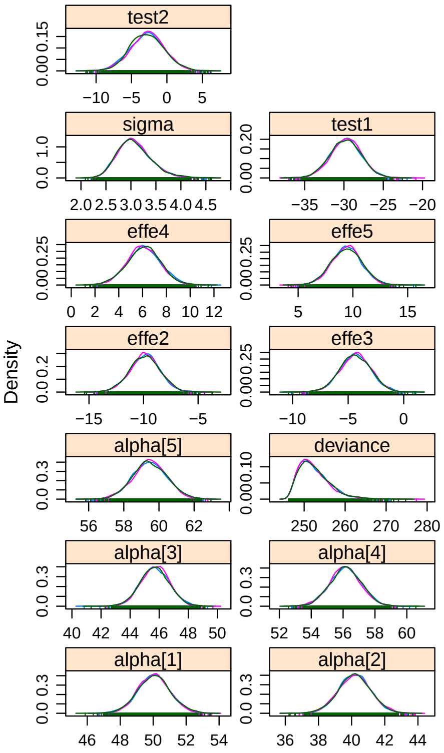

Figure 2.2: MCMC densities

3 Random effects ANOVA

The difference here is the we assume that population means come from a Gaussian distribution.

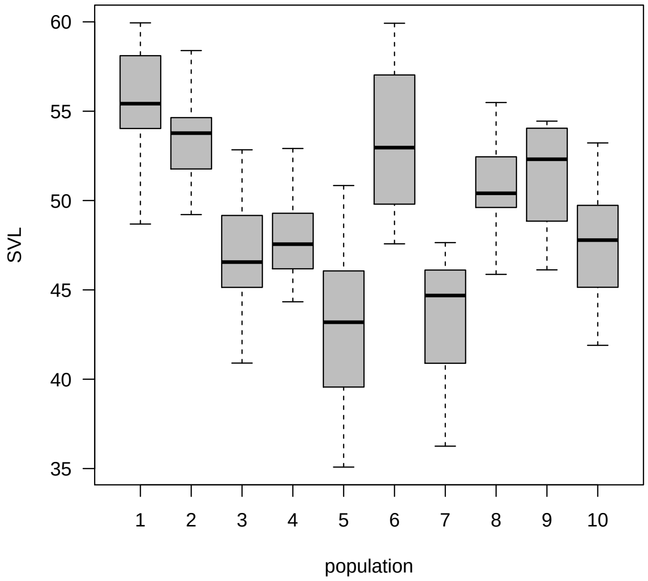

Figure 3.1: The distributions of the 10 populations of snakes with respect to snout-vent-length

3.1 Maximum likelihood analysis

```
library("lme4")
data = read.csv("re-svl.csv")
pop <- as.factor(data$x)
lme.fit <- lmer(data$y~1+1 | pop, REML=TRUE)
print(summary(lme.fit))
print(ranef(lme.fit))
print(summary(lme.fit)$coeff,dig=3)
print(paste(’AIC’, AIC(lme.fit)))
```

```
[1] "..."
Linear mixed model fit by REML [’lmerMod’]
Formula: data$y ~ 1 + 1 | pop

REML criterion at convergence: 655.1

Scaled residuals:
    Min 1Q Median 3Q Max
-2.48988 -0.71998 0.00715 0.67741 2.24779

Random effects:
 Groups Name Variance Std.Dev.
 pop (Intercept) 16.85 4.105
 Residual 11.06 3.326
Number of obs: 120, groups: pop, 10

Fixed effects:
           Estimate Std. Error t value
(Intercept) 49.362 1.333 37.03
$pop
   (Intercept)
1 5.728721
2 3.894882
3 -2.228275
4 -1.460284
5 -5.998548
6 3.706515
7 -5.559655
8 1.302983
9 2.067522
10 -1.453860

           Estimate Std. Error t value
(Intercept) 49.4 1.33 37
[1] "AIC 661.126107734498"
```

3.2 MCMC analysis

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 273

Initializing model

$BUGSoutput
Inference for Bugs model at "re-anova.txt", fit using jags,
 3 chains, each with 5000 iterations (first 200 discarded), n.thin = 2
 n.sims = 7200 iterations saved
             mean  sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
deviance     630.2 5.0 622.6 626.5 629.5 633.0 642.0    1  7200
effe[1]        5.9 1.8   2.5   4.7   5.9   7.1   9.5    1  7200
effe[2]        4.1 1.8   0.5   2.9   4.0   5.2   7.8    1  7200
effe[3]       -2.1 1.8  -5.6  -3.3  -2.1  -0.9   1.5    1  7200
effe[4]       -1.3 1.8  -4.9  -2.5  -1.3  -0.2   2.2    1  7200
effe[5]       -5.9 1.8  -9.5  -7.0  -5.9  -4.7  -2.3    1  7200
effe[6]        3.9 1.8   0.4   2.7   3.8   5.0   7.5    1  7200
effe[7]       -5.4 1.8  -9.0  -6.6  -5.4  -4.3  -1.9    1  7200
effe[8]        1.4 1.8  -2.1   0.3   1.4   2.6   5.0    1  7200
effe[9]        2.2 1.8  -1.4   1.1   2.2   3.3   5.8    1  7200
effe[10]      -1.4 1.8  -4.9  -2.5  -1.4  -0.2   2.2    1  7200
mu            49.2 1.6  46.0  48.2  49.3  50.2  52.4    1  7200
pop.mean[1]   55.1 1.0  53.2  54.5  55.1  55.8  57.0    1  6200
pop.mean[2]   53.3 1.0  51.4  52.7  53.3  53.9  55.1    1  7200
pop.mean[3]   47.1 0.9  45.3  46.5  47.1  47.8  49.0    1  7200
pop.mean[4]   47.9 0.9  46.0  47.3  47.9  48.5  49.8    1  7200
pop.mean[5]   43.3 1.0  41.5  42.7  43.3  44.0  45.2    1  1700
pop.mean[6]   53.1 0.9  51.2  52.5  53.1  53.7  55.0    1  7200
pop.mean[7]   43.8 1.0  41.9  43.1  43.8  44.4  45.7    1  7200
pop.mean[8]   50.7 0.9  48.8  50.0  50.6  51.3  52.5    1  7200
pop.mean[9]   51.4 1.0  49.6  50.8  51.4  52.1  53.3    1  7200
pop.mean[10]  47.9 1.0  46.0  47.2  47.9  48.5  49.8    1  7200
sigma.group    4.8 1.3   2.9   3.8   4.5   5.5   8.0    1  2100
sigma.res      3.4 0.2   3.0   3.2   3.4   3.5   3.8    1  7200

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 12.5 and DIC = 642.7
DIC is an estimate of expected predictive error (lower deviance is better).
```

The truth of the among population `svl` variation is 5. In the Bayesian analysis we look at `sigma.group` and for the REML estimate `pop (intercept)`.

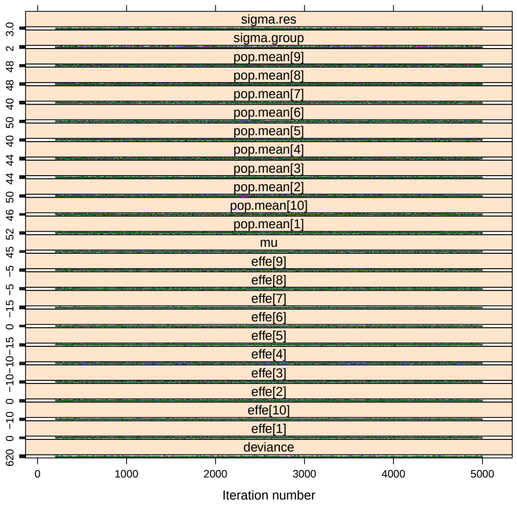

Figure 3.2: The MCMC chains

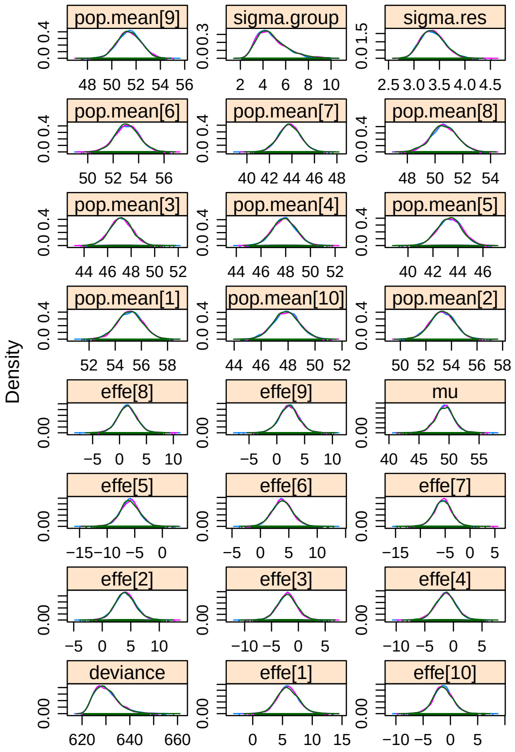

Figure 3.3: MCMC densities

4 Reproducibility

To reproduce this document.

- `$ Rscript generateData.R`
- `$ Rscript fixedEffectAnova.R > fe-anova-out.txt`
- `$ Rscript randomEffectAnova.R > re-anova-out.txt`
- `$ python run.py`

Bibliography

[1] M. Kery. *Introduction to WinBUGS for Ecologists*, Elsevier Academic Press, 2010.