Outline

1. Random and mixed effect models: Introduction
   - Corn example
   - Random effects for sites
   - Motivations for random and mixed effect models
   - Partial pooling to estimate random effects
   - Nested effects: Flowering time example
   - Mixing random and fixed effects

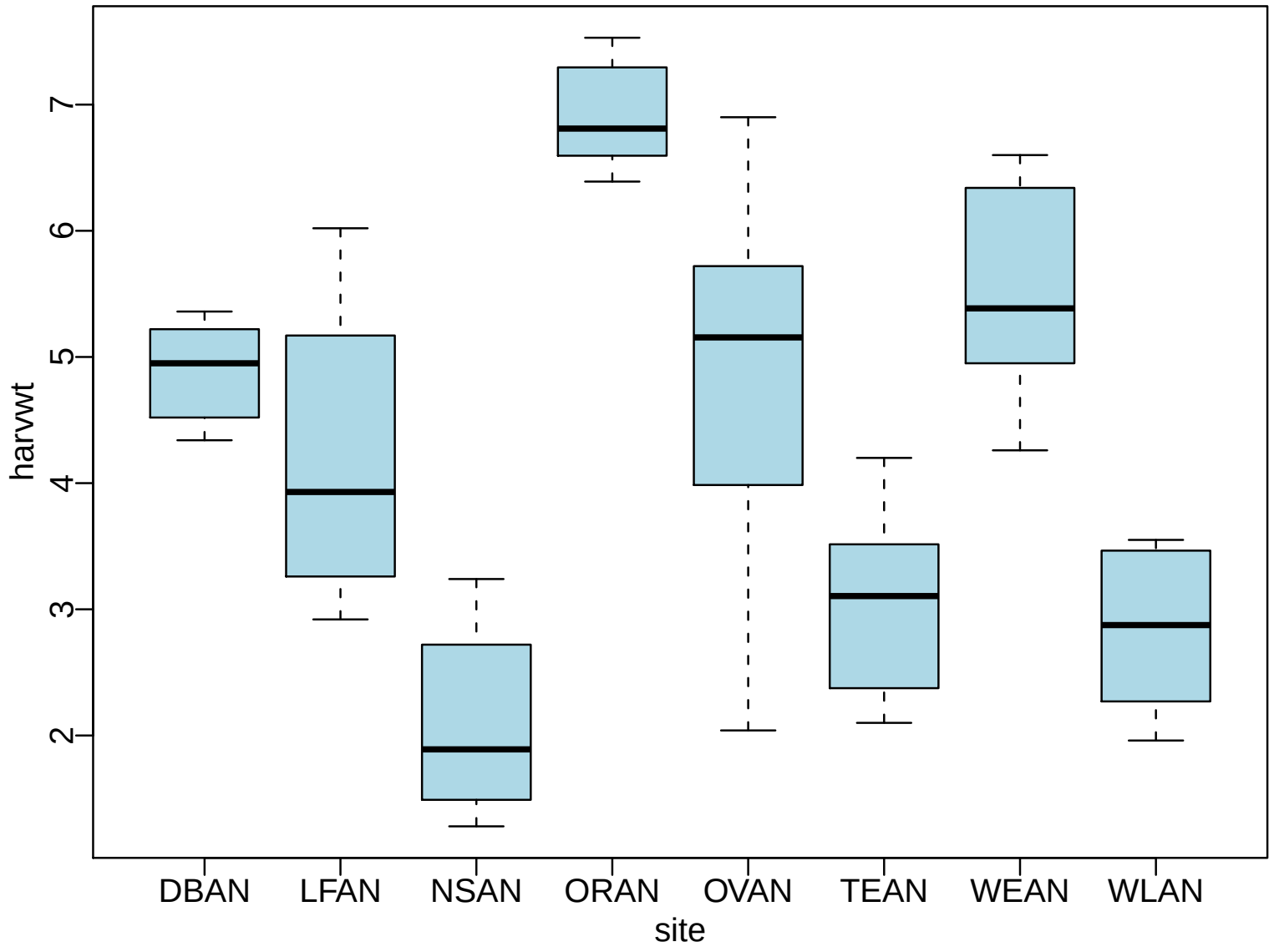
Corn example

- Subset of a larger data set on corn grown on the island Antigua.
- Response: harvest weight (`harvwt`) per plot.
- 8 sites, with 8 separate plots within each site where the corn is grown under the same treatment conditions.
- Does the site have an effect on the harvest weight?

```
> corn = read.table("corn.txt", header = T)
> summary(corn)
      site    block          ears           harvwt
 DBAN   : 8   I  :16   Min.   :13.00   Min.   :1.280
 LFAN   : 8   II :16   1st Qu.:37.75   1st Qu.:2.935
 NSAN   : 8   III:16   Median :43.00   Median :4.300
 ORAN   : 8   IV :16   Mean   :41.22   Mean   :4.292
 OVAN   : 8            3rd Qu.:46.00   3rd Qu.:5.442
 TEAN   : 8            Max.   :58.00   Max.   :7.530
 (Other):16
```

Corn example

```
> plot(harvwt ~ site,data=corn,pch=16,col="lightblue")
```



Corn example

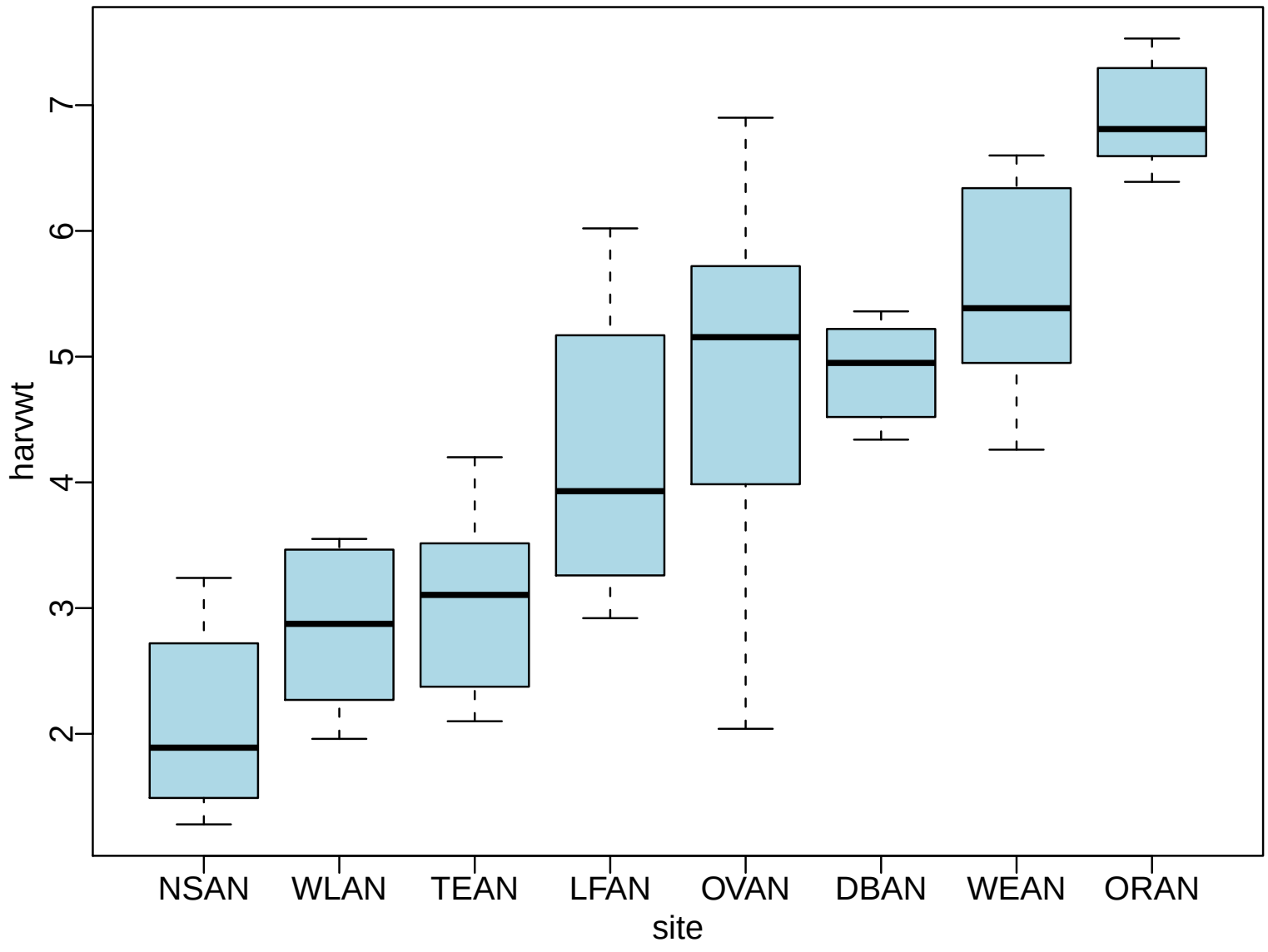
```
> corn$site
 [1] DBAN DBAN DBAN DBAN DBAN DBAN DBAN DBAN LFAN LFAN LFAN LFAN LFAN
[16] LFAN TEAN TEAN TEAN TEAN TEAN TEAN TEAN TEAN WEAN WEAN WEAN WEAN
...
[61] ORAN ORAN ORAN ORAN
Levels: DBAN LFAN NSAN ORAN OVAN TEAN WEAN WLAN

# re-ordering sites, from lowest to highest mean harvest:

> corn$site = with(corn, reorder(site, harvwt))
> corn$site
 [1] DBAN DBAN DBAN DBAN DBAN DBAN DBAN DBAN LFAN LFAN LFAN LFAN LFAN
[16] LFAN TEAN TEAN TEAN TEAN TEAN TEAN TEAN TEAN WEAN WEAN WEAN WEAN
...
[61] ORAN ORAN ORAN ORAN
attr(,"scores")
   DBAN    LFAN    NSAN    ORAN    OVAN    TEAN    WEAN    WLAN
4.88500 4.20750 2.09000 6.91500 4.83250 3.03625 5.52625 2.84125
Levels: NSAN WLAN TEAN LFAN OVAN DBAN WEAN ORAN

# then plot again:
> plot(harvwt ~ site,data=corn,pch=16,col="lightblue")
```

Corn example



First analysis: Standard one-way ANOVA

Standard regression: we could use one-way ANOVA with 8 fixed parameters for the site mean weights and a single plot-level source of error.

Model:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 \cdot \mathbb{1}_{\text{site 2}} + \cdots + \beta_8 \cdot \mathbb{1}_{\text{site 8}} + e_i \\ &= \mu + \alpha_1 \cdot \mathbb{1}_{\text{site 1}} + \alpha_2 \cdot \mathbb{1}_{\text{site 2}} \cdots + \alpha_8 \cdot \mathbb{1}_{\text{site 8}} + e_i \end{aligned}$$

where $i = 1, \ldots, 64$ indexes observations and $e_i \sim \text{iid}\mathcal{N}(0, \sigma^2)$.

Fixed, unknown parameters:

- β_j, $j = 1, \ldots, 8$: intercept $\beta_1 =$ mean corn yield for site 1, and adjustments $\beta_2, \ldots, \beta_8$ for other sites,
  or μ and α_j, $j = 1, \ldots, 8$: intercept $\mu =$ grand mean over all sites, and adjustements $\alpha_1, \ldots, \alpha_8$ around this overall mean for each site, with the constraint $\sum_{j=1}^{8} \alpha_j = 0$.
- σ^2

Standard regression model

By default, R uses the β formula, with the first site as the reference level.

```
> corn.lm = lm(harvwt ~ site, data = corn)
> summary(corn.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.090      0.309    6.76  8.6e-09 ***
siteWLAN       0.751      0.437    1.72    0.091 .
siteTEAN       0.946      0.437    2.16    0.035 *
siteLFAN       2.118      0.437    4.84  1.0e-05 ***
siteOVAN       2.743      0.437    6.27  5.5e-08 ***
siteDBAN       2.795      0.437    6.39  3.5e-08 ***
siteWEAN       3.436      0.437    7.86  1.3e-10 ***
siteORAN       4.825      0.437   11.03  1.1e-15 ***

Residual standard error: 0.88 on 56 degrees of freedom
Multiple R-squared: 0.767,      Adjusted R-squared: 0.737
F-statistic: 26.3 on 7 and 56 DF,  p-value: 1.55e-15
```

Standard regression model

We may request the formula with an overall mean and α adjustments that sum up to 0:

```
> options(contrasts=c("contr.sum", "contr.poly"))
> corn.lm = lm(harvwt ~ site, data = corn)
> summary(corn.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.2917     0.1093   39.26  < 2e-16 ***
site1        -2.2017     0.2892   -7.61  3.4e-10 ***
site2        -1.4505     0.2892   -5.02  5.7e-06 ***
site3        -1.2555     0.2892   -4.34  6.0e-05 ***
site4        -0.0842     0.2892   -0.29    0.772
site5         0.5408     0.2892    1.87    0.067 .
site6         0.5933     0.2892    2.05    0.045 *
site7         1.2345     0.2892    4.27  7.7e-05 ***

Residual standard error: 0.875 on 56 degrees of freedom
Multiple R-squared: 0.767,      Adjusted R-squared: 0.737
F-statistic: 26.3 on 7 and 56 DF,  p-value: 1.55e-15
```

Site means

```
> means     = with(corn, tapply(harvwt, site, mean))
> means
  NSAN   WLAN   TEAN   LFAN   OVAN   DBAN   WEAN   ORAN
2.0900 2.8412 3.0362 4.2075 4.8325 4.8850 5.5262 6.9150

> mean(means)
[1] 4.291719
```

The adjustment α_8 for the last site was not given, but it has to be $-(-2.2017 - 1.4505 + \cdots + 1.2345)$.

Second analysis: random effect model

or one-way ANOVA with random effects.

We can consider the 8 sites as part of a larger population of sites across the island, and consider their mean harvest weights as random from some normal distribution.

Second analysis: random effect model

Model:

$$y_i = \mu + \alpha_{j[i]} + e_i$$

where $j[i] = 1, \ldots, 8$ indicates which of the 8 sites contains the ith observation.

Two levels of variation:

- at the plot level: $e_i \sim$ iid $\mathcal{N}(0, \sigma^2)$.
- at the site level: $\alpha_j \sim$ iid $\mathcal{N}(0, \sigma^2_\alpha)$ are random effects for the sites, $j = 1, \ldots, 8$.

Fixed, unknown parameters:

- μ_α: overall mean corn yield over the entire population of plots and sites.
- σ^2_α: variance of the mean sites' corn yield over the population of sites.
- and σ^2: variance of the plot corn yield over the population of plots within the same site.

Using `lmer` in R

`lme4`: most recently developed R package for fitting linear models with random effects. To install it:

```
>install.packages("lme4")
```

`lmer`: function to use instead of `lm` (Linear Mixed Effects in R). To use it, first load the package, once per session:

```
>library(lme4)
```

The model formula with a random effect in `lmer` differs from `lm`: need to include a term of the form `(pred | group)` where

- `group` is the variable defining the groups within which the random effect applies.
- `pred` defines a model matrix (often just an intercept 1): each group is to have its own coefficients for this regression model.

Random effect model with `lmer`

```
> corn.lmer = lmer(harvwt ~ (1 | site), data = corn)
> summary(corn.lmer)

Linear mixed model fit by REML

Formula: harvwt ~ (1 | site)
   Data: corn
 AIC BIC logLik deviance REMLdev
 195 201  -94.5      190     189

Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.417    1.555
 Residual             0.765    0.875
Number of obs: 64, groups: site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)     4.29       0.56    7.66
```

Comparing models

- Same residual $\hat{\sigma} = 0.87$ ($\hat{\sigma}^2 = 0.765$): plot-level variation
- Same overall mean value: 4.29

```
> means     = with(corn, tapply(harvwt, site, mean))
> means
  NSAN   WLAN   TEAN   LFAN   OVAN   DBAN   WEAN   ORAN
2.0900 2.8412 3.0362 4.2075 4.8325 4.8850 5.5262 6.9150
> mean(means)
[1] 4.291719
```

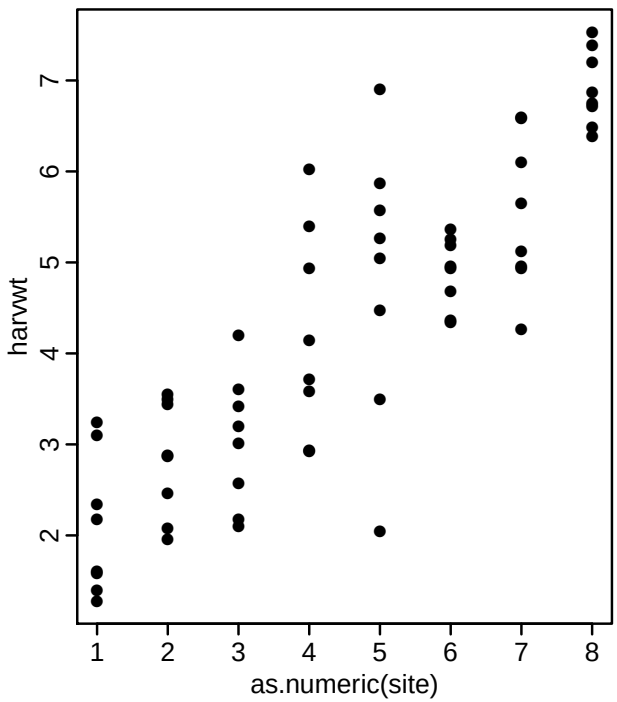
- Estimated variance of observations: $1.55^2 + 0.87^2$
  Covariance of observations at the same site: 1.55^2
  Square correlation of observations at the same site: $1.55^2/(1.55^2 + 0.87^2) = 0.76$
- This is close to $R^2 = 0.77$ in standard one-way ANOVA, which corresponds to how much variance is explained by the site-level variability.

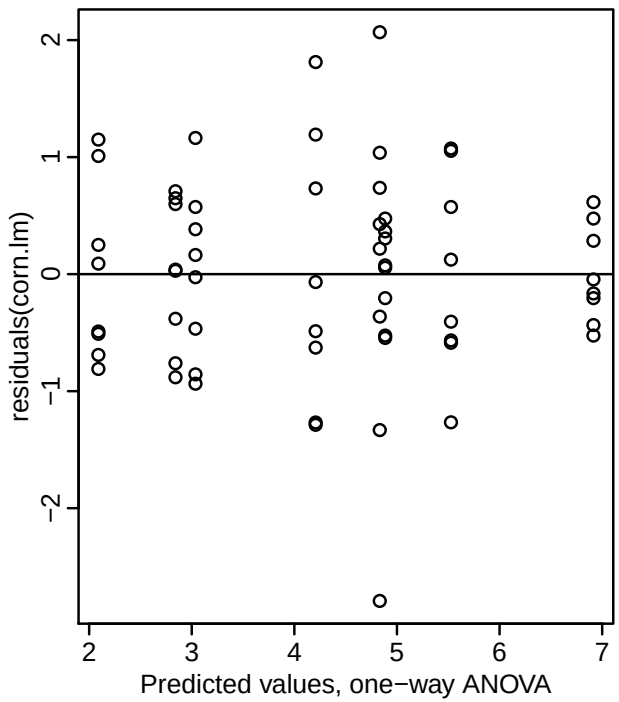
Residual plot

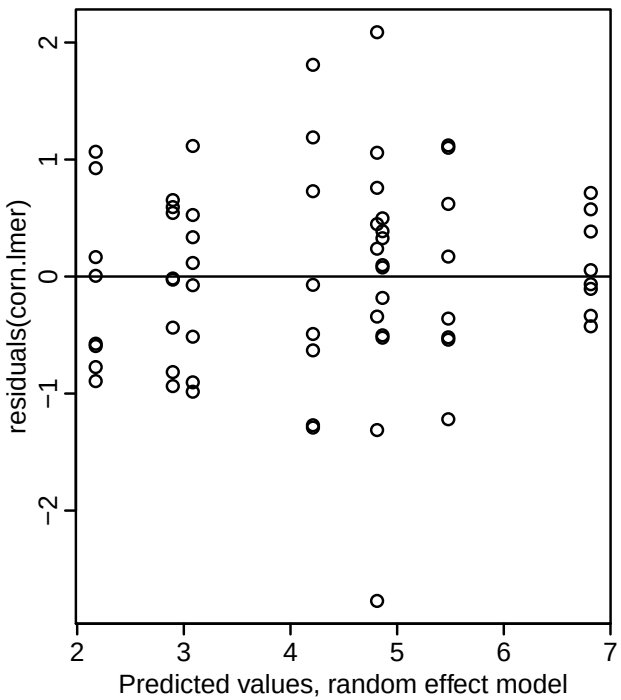
```
layout(matrix(1:3, 1,3))
plot(harvwt ~ as.numeric(site),data=corn,pch=16) # plot data first

plot(residuals(corn.lm) ~ fitted(corn.lm),        #residual plot from
     xlab="Predicted values, one-way ANOVA")      #standard 1-way ANOVA
abline(h=0)

plot(residuals(corn.lmer) ~ fitted(corn.lmer),   #from random effect
    xlab="Predicted values, random effect model")              # model
abline(h=0)
```







Motivations for random effect models

Random effects are used to include sources of variation at more than one level. Examples:

- Repeated measures — when a single individual is measured multiple times, it is often appropriate to model two levels of variation, one for individuals and one for measurements.
- Split-plot designs — in agricultural or ecological studies, it is often the case that sites are broken into plots and possibly subplots. Variables can be measured at the site, plot, subplot, or individual measurement level.
- Also appropriate for non-nested variables. For example, measurements could be clustered by year and by site if a single site is measured over multiple years.

Motivations for random effect models

- Accounting for both individual and group level variation in estimating group-level effects.
- Modeling individual level regression coefficients.
- Estimation of effects for subgroups.

We can have covariates and error associated with the plot level and separate covariates and error associated with the site level.

Partial pooling to estimate random effects

Corn yield:

$$y_i = \mu + \alpha_{j[i]} + e_i$$

where $i = 1, \ldots, 64$ indexes the observation and $j[i] = 1, \ldots, 8$ indicates which of the 8 sites contains the ith observation.

- $\alpha_j, \sim \mathcal{N}(0, \sigma_\alpha^2)$ are random coefficients: we do not directly try to estimate them;
- $e_i \sim \text{iid}\mathcal{N}(0, \sigma^2)$ are random, as usual.
- μ, σ_α^2, and σ^2 are fixed parameters. We do want to estimate them, by finding the numerical values that makes the likelihood $L(\mu, \sigma_\alpha^2, \sigma^2)$ as large as possible.

After the fact: after we have estimated $\hat{\mu}$, $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}^2$, we may estimate site means $\hat{\alpha}_j$ just like the residuals $r_i = y_i - \hat{y}_i$ are estimates of the 'errors' e_i.

We will compare the estimated sites effects $\hat{\alpha}_j$ when they are considered fixed and when they are considered random:

Compare fitted values $\mu + \alpha_j$

```
> means      = with(corn, tapply(harvwt, site, mean))
> mean(means)
[1] 4.291719
> fitted.lm  = with(corn, tapply(fitted(corn.lm), site, mean))
> fitted.lmer= with(corn, tapply(fitted(corn.lmer),site,mean))
> round( rbind(means,fitted.lm,fitted.lmer), 3)

             DBAN LFAN NSAN ORAN OVAN TEAN WEAN WLAN
means       4.88 4.21 2.09 6.92 4.83 3.04 5.53 2.84
fitted.lm   4.88 4.21 2.09 6.92 4.83 3.04 5.53 2.84
fitted.lmer 4.86 4.21 2.17 6.82 4.81 3.08 5.48 2.90
```

The random effect model shrinks the estimated site means toward the overall mean.

Likelihood

- We can express the likelihood of the data in terms of the random effects and the fixed parameters.
- The normal density $f(x|\mu,\sigma)$ for mean μ and standard deviation σ is large when the x is close to μ.
- The likelihood has the form

$$L = \left(\prod_{j=1}^{J} f(\mu+\alpha_j|\mu,\sigma_\alpha)\right)\left(\prod_{i=1}^{n} f(y_i|\mu+\alpha_{j[i]},\sigma)\right)$$

  left factor: large when all of the $\mu+\alpha_j$ are close to μ_α.
  right factor: large when $\mu+\alpha_j$ are close to the y_i in group j.
- The best estimate of μ will be the overall mean $\bar{y}_{\text{all}}$.
- The likelihood is maximized at a balancing point when the $\mu+\alpha_j$ is somewhere between the overall mean $\bar{y}_{\text{all}}$ and the sample means $\bar{y}_j$.

Pooling

- Complete pooling: assumes no difference between groups:

$$\mu + \alpha_1 = \cdots = \mu + \alpha_J = \bar{y}_{\text{all}}$$

  extreme case model where $\sigma_\alpha = 0$.
- No pooling: estimates each α_j using only data from the jth sample:

$$\mu + \alpha_1 = \bar{y}_1, \ \ldots, \ \mu + \alpha_J = \bar{y}_J$$

  extreme case where $\sigma_\alpha = +\infty$.
- Random effect models correspond to partial pooling: data from other samples affect estimates for sample j, but the data within sample j is most influential.
- Estimated coefficients: weighted average between the corresponding sample mean and the grand mean.

$$\mu + \hat{\alpha}_j \approx \frac{\left(\frac{n_j}{\sigma^2}\right)\bar{y}_j + \left(\frac{1}{\sigma_\alpha^2}\right)\bar{y}_{\text{all}}}{\frac{n_j}{\sigma^2} + \frac{1}{\sigma_\alpha^2}}$$

Weighted average

$$\mu + \hat{\alpha}_j \approx \frac{\left(\frac{n_j}{\sigma^2}\right)\bar{y}_j + \left(\frac{1}{\sigma_\alpha^2}\right)\bar{y}_{\text{all}}}{\frac{n_j}{\sigma^2} + \frac{1}{\sigma_\alpha^2}} = \left(\frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\sigma_\alpha^2}}\right)\bar{y}_j + \left(\frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\sigma_\alpha^2}}\right)\bar{y}_{\text{all}}$$

- An expression of the form

  $$w_1 A + w_2 B$$

  where $w_1 + w_2 = 1$ is a weighted average of A and B.
- The relative distance of the average to the endpoints is inversely proportional to the weights.

  (If one weight is ten times the other, the distance of the average to the end with the higher weight will be ten times smaller than to the other.)

Interpretations

$$\mu + \hat{\alpha}_j \approx \frac{\left(\frac{n_j}{\sigma^2}\right)\bar{y}_j + \left(\frac{1}{\sigma^2_\alpha}\right)\bar{y}_{\text{all}}}{\frac{n_j}{\sigma^2} + \frac{1}{\sigma^2_\alpha}}$$

$\mu + \hat{\alpha}_j$ moves closer to $\bar{y}_j$ when:

- n_j gets bigger (more direct data in the group), or;
- σ^2 gets smaller (more likely for y_i to be close to group means)

$\mu + \hat{\alpha}_j$ moves closer to $\bar{y}_{\text{all}}$ when

- σ^2_α gets smaller (more likely for all α_j to be close to each other)

Numerical example

$n_j = 8$ for $j = 1, \ldots, 8$, $\quad \hat{\mu} = \bar{y}_{\text{all}} = 4.29$, $\quad \hat{\sigma}_\alpha = 1.55$, $\quad \hat{\sigma} = 0.87$

Site DBAN:

$\bar{y}_1 = 4.88$, $\hat{\mu} + \hat{\alpha}_1 = 4.86$

$\frac{8}{(0.87)^2} = 10.57$

$\frac{1}{(1.55)^2} = 0.42$

$$\frac{10.57}{10.99}(4.88) + \frac{0.42}{10.99}(4.29) = 4.86$$

The estimate is shrunk very little toward the grand mean.

Site NSAN:

$\bar{y}_3 = 2.09$, $\hat{\mu} + \hat{\alpha}_3 = 2.17$

$\frac{8}{(0.87)^2} = 10.57$

$\frac{1}{(1.55)^2} = 0.42$

$$\frac{10.57}{10.99}(2.09) + \frac{0.42}{10.99}(4.29) = 2.17$$

The estimate is shrunk a bit more toward the grand mean: the same proportion of a longer distance is larger.

Nested effects: Flowering time example

Flowering time: trait that breeders are often interested in. What genes are controlling it? How much variability is there in flowering time *between* subspecies of Brassica rapa, and between plants *within* subspecies?

`dtf`: date to flowering (days). If $>$ experiment duration ($\sim$ 90 days), then NA. Some subspecies needed vernalization.
`subspecies`: predictor of interest.
`inventoryID`: packet from which 4 plants were measured.
`rep`: from 1 to 4. Identifies the plant # from a specific packet.
`exptID`: random number used for blinding and randomization.

Flowering time

```
> brassica = read.table("floweringTime.txt", header=T)

> brassica$rep = as.factor(brassica$rep) # rep should be categorical

> brassica$logdtf = log(brassica$dtf)    # useful later

> str(brassica)
'data.frame':   188 obs. of  6 variables:
 $ exptID     : int  26 82 106 144 4 67 158 177 6 45 ...
 $ inventoryID: Factor w/ 47 levels "EF061","EF064",..: 1 1 1 1 2 2 2
 $ subspecies : Factor w/ 11 levels "chinensis","dichotoma",..: 9 9 9
 $ rep        : Factor w/ 4 levels "1","2","3","4": 1 4 2 3 2 3 1 4 2
 $ dtf        : int  NA NA NA NA NA NA NA NA 44 48 ...
 $ logdtf     : num  NA NA NA NA NA ...
```

Flowering time: dealing with missing data

```
> summary(brassica)

     exptID       inventoryID          subspecies rep
 Min.   :  1.0   EF061  :  4   chinensis   :32   1:47
 1st Qu.: 47.7   EF064  :  4   trilocularis:32   2:47
 Median : 94.5   EF076  :  4   dichotoma   :16   3:47
 Mean   : 94.5   EF076/2:  4   oleifera    :16   4:47
 3rd Qu.:141.2   EF077  :  4   rapifera    :16
 Max.   :188.0   EF089  :  4   (Other)     :56
                 (Other):164   NA's        :20

      dtf           logdtf
 Min.   :25     Min.   : 3.219
 1st Qu.:41     1st Qu.: 3.714
 Median :49     Median : 3.892
 Mean   :54.84  Mean   : 3.952
 3rd Qu.:70     3rd Qu.: 4.248
 Max.   :93     Max.   : 4.533
 NA's   :38     NA's   :38.000

> levels(brassica$subspecies)
 [1] "chinensis"     "dichotoma"     "narinosa"      "nipposonica"
 [5] "oleifera"      "parachinensis" "pekinensis"    "perviridis"
 [9] "rapifera"      "trilocularis"  "utilis"
```

Flowering time: dealing with missing data

```
> brassica[1:18,]
   exptID inventoryID subspecies rep dtf   logdtf
1      26       EF061   rapifera   1  NA       NA
2      82       EF061   rapifera   4  NA       NA
3     106       EF061   rapifera   2  NA       NA
4     144       EF061   rapifera   3  NA       NA
5       4       EF064   rapifera   2  NA       NA
6      67       EF064   rapifera   3  NA       NA
7     158       EF064   rapifera   1  NA       NA
8     177       EF064   rapifera   4  NA       NA
9       6       EF076     utilis   2  44 3.784190
10     45       EF076     utilis   3  48 3.871201
11    111       EF076     utilis   4  56 4.025352
12    180       EF076     utilis   1  42 3.737670
13     61     EF076/2     utilis   2  49 3.891820
14     80     EF076/2     utilis   3  46 3.828641
15    159     EF076/2     utilis   1  48 3.871201
16    175     EF076/2     utilis   4  44 3.784190
17     52       EF077       <NA>   4  33 3.496508
18     73       EF077       <NA>   3  32 3.465736
> is.na(brassica$dtf)
  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALS
 [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALS
 [25]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
...
```

Flowering time: dealing with missing `dtf`

Missing data: non-observed `dtf`, unknown subspecies name.

```
> with(brassica, table(subspecies)) # some subsp. sampled more
subspecies
    chinensis    dichotoma     narinosa   nipposonica      oleifera
           32           16            8             8            16
parachinensis   pekinensis   perviridis      rapifera  trilocularis
            4           12            8            16            32
       utilis
           16
> brassica2 =subset(brassica, !is.na(dtf)) # subset to non-missing dft

> with(brassica2, table(subspecies))
subspecies
    chinensis    dichotoma     narinosa   nipposonica      oleifera
           32           16            5             0            16
parachinensis   pekinensis   perviridis      rapifera  trilocularis
            4            8            6             0            32
       utilis
           14
# some subspecies are no longer represented. vernalization!
# To remove unrepresented levels of subspecies:
> brassica2$subspecies = factor(brassica2$subspecies)
```

Flowering time: dealing with missing subspecies

```
> with(brassica2, table(subspecies))
subspecies
    chinensis     dichotoma      narinosa      oleifera parachinensis
           32            16             5            16             4
   pekinensis    perviridis  trilocularis        utilis
            8             6            32            14
> brassica2[1:15,]  # some unknown subspecies.
   exptID inventoryID subspecies rep dtf   logdtf
9       6       EF076     utilis   2  44 3.784190
10     45       EF076     utilis   3  48 3.871201
11    111       EF076     utilis   4  56 4.025352
12    180       EF076     utilis   1  42 3.737670
13     61     EF076/2     utilis   2  49 3.891820
14     80     EF076/2     utilis   3  46 3.828641
15    159     EF076/2     utilis   1  48 3.871201
16    175     EF076/2     utilis   4  44 3.784190
17     52       EF077       <NA>   4  33 3.496508
18     73       EF077       <NA>   3  32 3.465736
19    143       EF077       <NA>   2  35 3.555348
20    157       EF077       <NA>   1  39 3.663562
21     18       EF089       <NA>   4  65 4.174387
23    118       EF089       <NA>   2  63 4.143135
24    149       EF089       <NA>   1  68 4.219508
# Remove samples:
> brassica2 = subset(brassica2, !is.na(subspecies))
```

Flowering time: dealing with missing data

```
> with(brassica2, table(inventoryID))
inventoryID
  EF061   EF064   EF076 EF076/2   EF077   EF089   EF090   EF091   EF09
      0       0       4       4       0       0       0       0
...
  EF148   EF149   EF150   EF151   EF153   EF154   EF156
      1       0       4       3       3       4       3
# some packets are no longer represented.
# So we remove those unrepresented packet levels:
> brassica2$inventoryID = factor(brassica2$inventoryID)
> with(brassica2, table(inventoryID)) # most packs have 4 reps
inventoryID
  EF076 EF076/2   EF092   EF093   EF094   EF102   EF103   EF104   EF10
      4       4       4       4       4       2       4       4
...
  EF148   EF150   EF151   EF153   EF154   EF156
      1       4       3       3       4       3
```

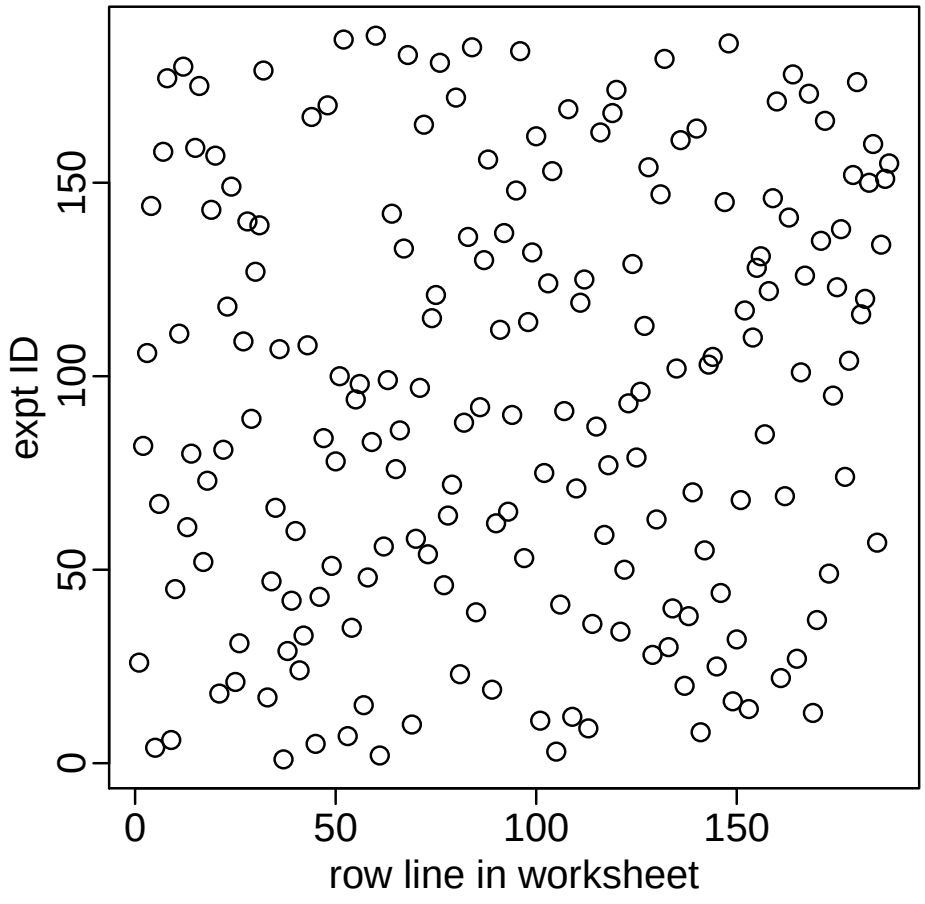
Flowering time: dealing with missing data

```
> brassica2[1:15,]   # Check that no more missing data
   exptID inventoryID subspecies rep dtf   logdtf
9       6       EF076     utilis   2  44 3.784190
10     45       EF076     utilis   3  48 3.871201
11    111       EF076     utilis   4  56 4.025352
12    180       EF076     utilis   1  42 3.737670
13     61     EF076/2     utilis   2  49 3.891820
14     80     EF076/2     utilis   3  46 3.828641
15    159     EF076/2     utilis   1  48 3.871201
16    175     EF076/2     utilis   4  44 3.784190
33     17       EF092  chinensis   2  82 4.406719
34     47       EF092  chinensis   1  87 4.465908
35     66       EF092  chinensis   3  70 4.248495
36    107       EF092  chinensis   4  59 4.077537
37      1       EF093  chinensis   1  70 4.248495
38     29       EF093  chinensis   2  60 4.094345
39     42       EF093  chinensis   4  93 4.532599
```

Flowering time: almost ready for analysis

- The selected data set is not completely balanced
- Blinding is important, and helps good randomization (plant placement in the greenhouse here).

```
> plot(brassica$exptID, xlab="row line in worksheet", ylab="expt ID")
```



Flowering time: Brainstorming for analysis

- What factors might affect flowering time `dtf`? What should we include as predictors?

```
> str(brassica2)
'data.frame':   133 obs. of  6 variables:
 $ exptID     : int  6 45 111 180 61 80 159 175 17 47 ...
 $ inventoryID: Factor w/ 36 levels "EF076","EF076/2",..: 1 1 1 1
 $ subspecies : Factor w/ 9 levels "chinensis","dichotoma",..: 9 9
 $ rep        : Factor w/ 4 levels "1","2","3","4": 2 3 4 1 2 3 1
 $ dtf        : int  44 48 56 42 49 46 48 44 82 87 ...
 $ logdtf     : num  3.78 3.87 4.03 3.74 3.89 ...
```

- Inventory ID is subspecies, rep is inventory ID.

Flowering time: Random effect model

We consider 3 levels of randomness:

- the 11 subspecies are part of a large population of Brassica rapa subspecies,
- within each subspecies, the seed packets are part of a large population of seed packets,
- within each packets, the 4 replicates are part of a large population of viable seeds.

Flowering time: Random effect model

Model:

$$y_i = \beta_0 + \gamma_{k[j[i]]} + \alpha_{j[i]} + e_i$$

where i is the plant (`rep`), $j[i]$ indicates the packet that i belongs to, $k[j]$ indicates the subspecies that packet j belongs to.

Random effects:

- $\gamma_k \sim \mathcal{N}(0, \sigma^2_\gamma)$, at the subspecies level,
- $\alpha_j \sim \mathcal{N}(0, \sigma^2_\alpha)$, at the packet level,
- $e_i \sim \mathcal{N}(0, \sigma^2_e)$, at the replicate level.

Fixed parameters:

- β_0 intercept: average flowering time across Brassica rapa subspecies
- σ^2_γ, σ^2_α, σ^2_e.

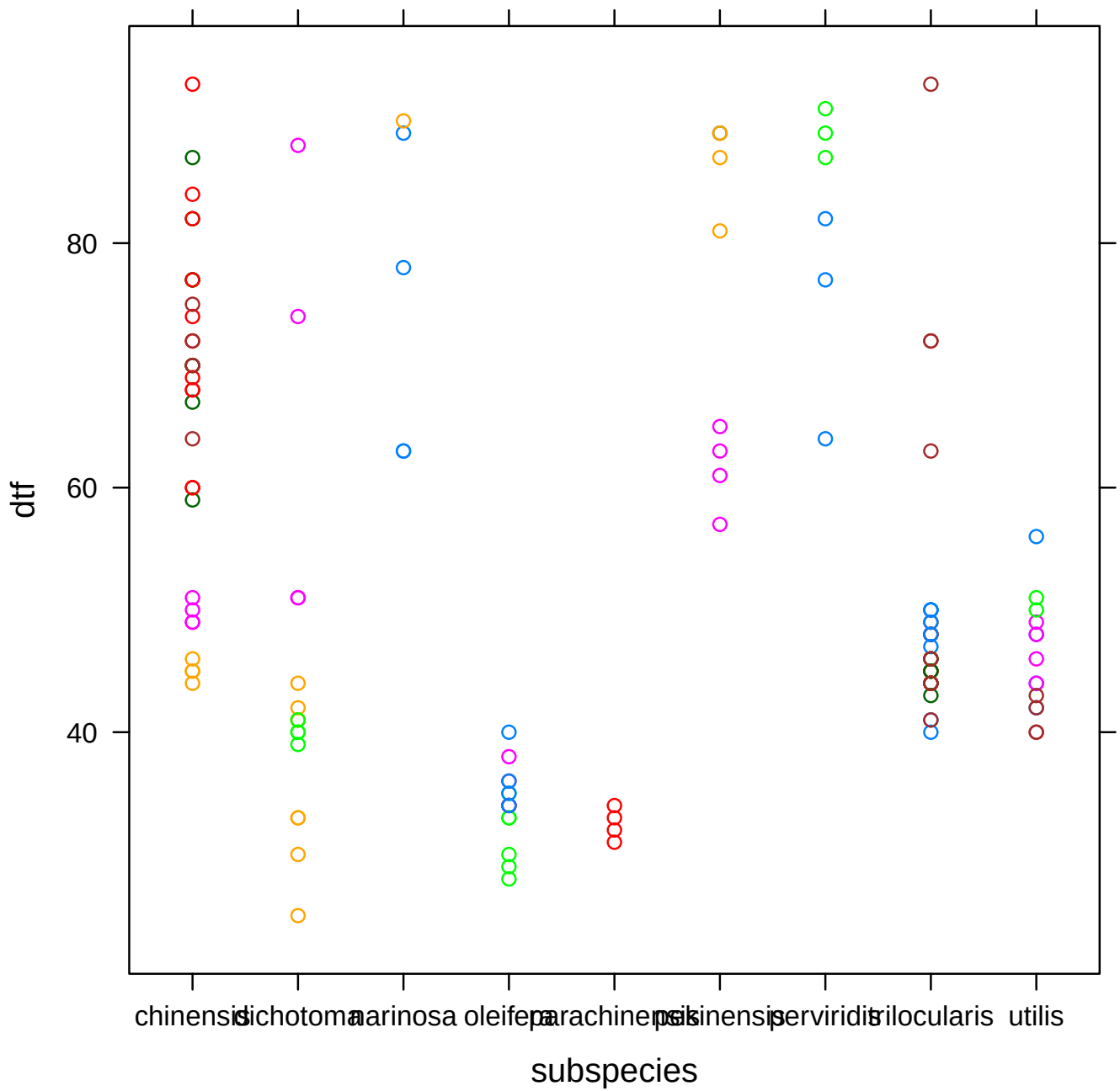
Flowering time: Plots

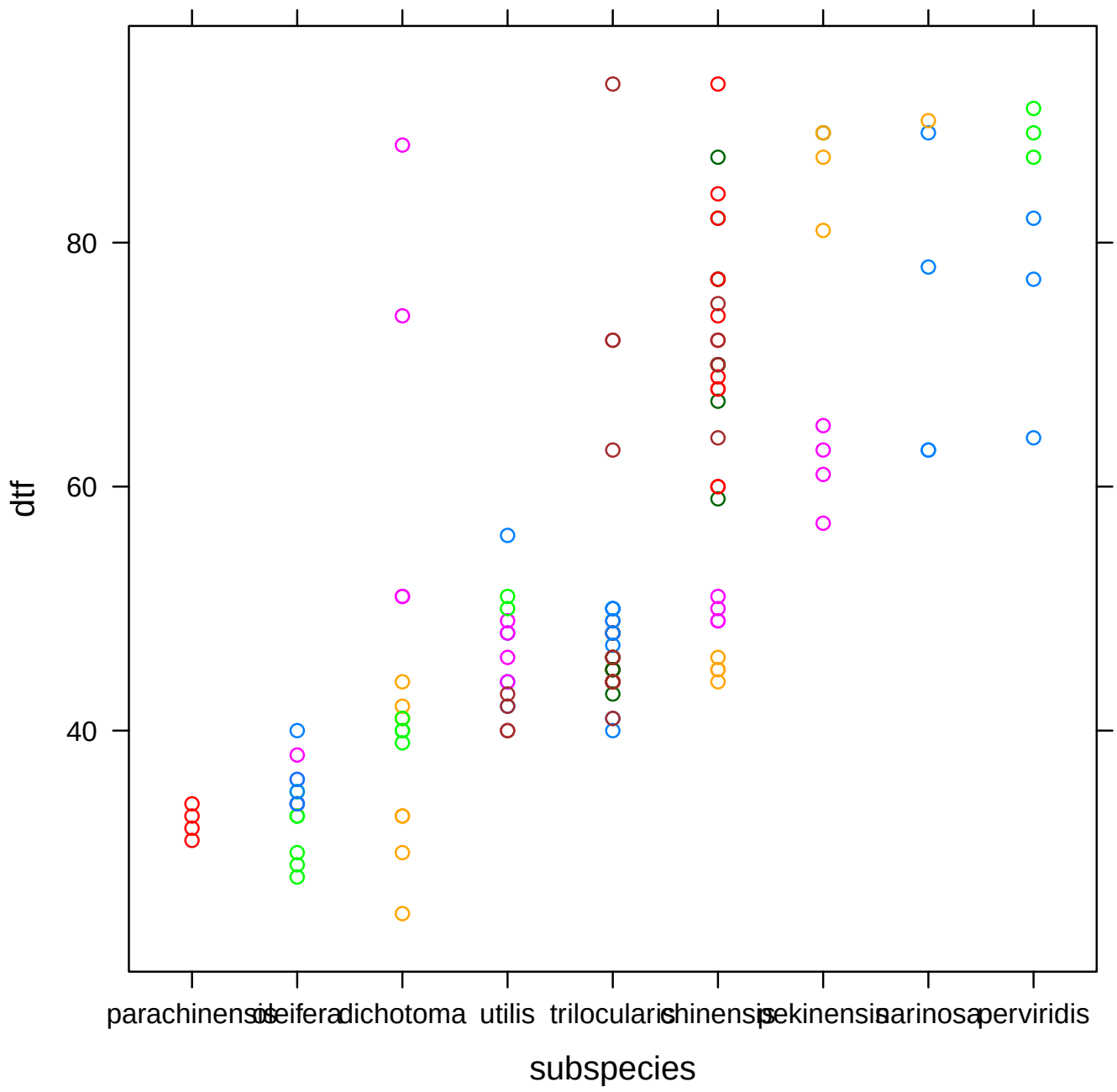
```
library(lattice)

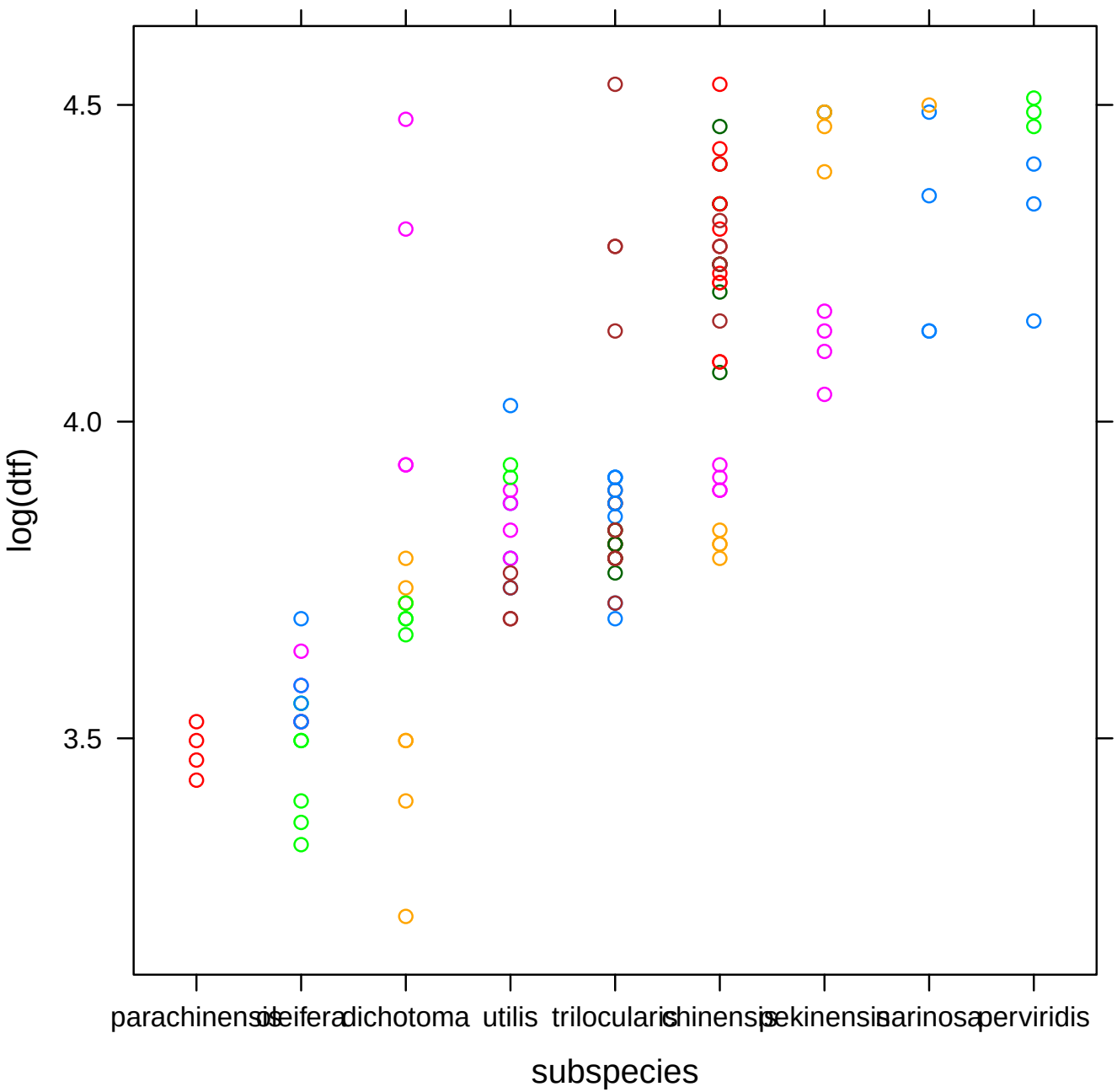
#--- dtf by subspecies, and we try a log transformation ----
xyplot(dtf ~ subspecies, group=inventoryID, brassica2)
brassica2$subspecies = reorder(brassica2$subspecies,
                               brassica2$dtf)
# re-order subspecies by increasing mean dtf
xyplot(  dtf ~ subspecies, group=inventoryID, data=brassica2)
xyplot(logdtf~ subspecies, group=inventoryID, data=brassica2)

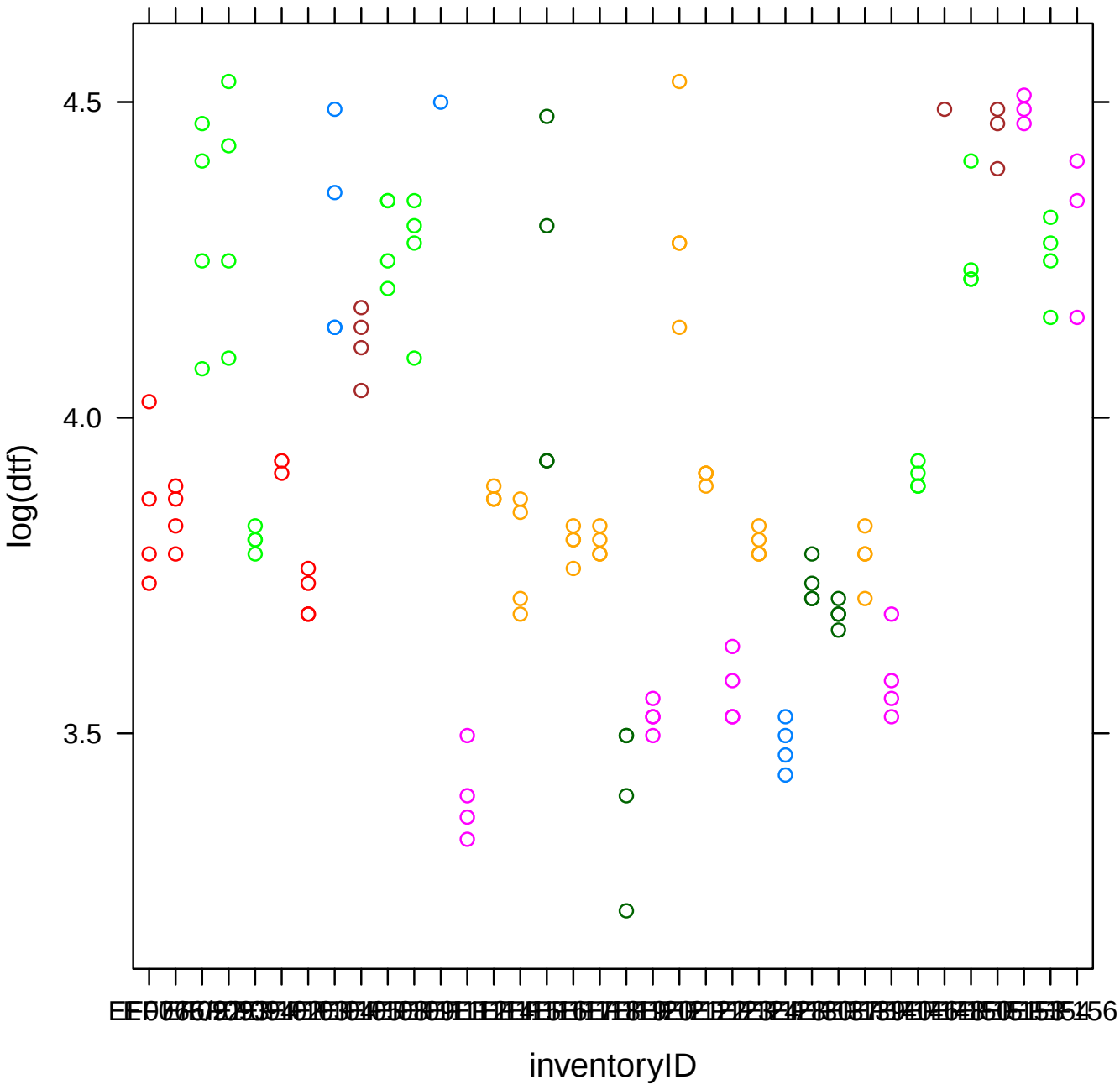
#--- dtf by packet -----------------------------------
xyplot(logdtf~ inventoryID, group=subspecies, data=brassica2)
brassica2$inventoryID = reorder(brassica2$inventoryID,
                                brassica2$dtf)
# re-order packets by increasing mean dtf
xyplot(logdtf~ inventoryID, group=subspecies, data=brassica2)

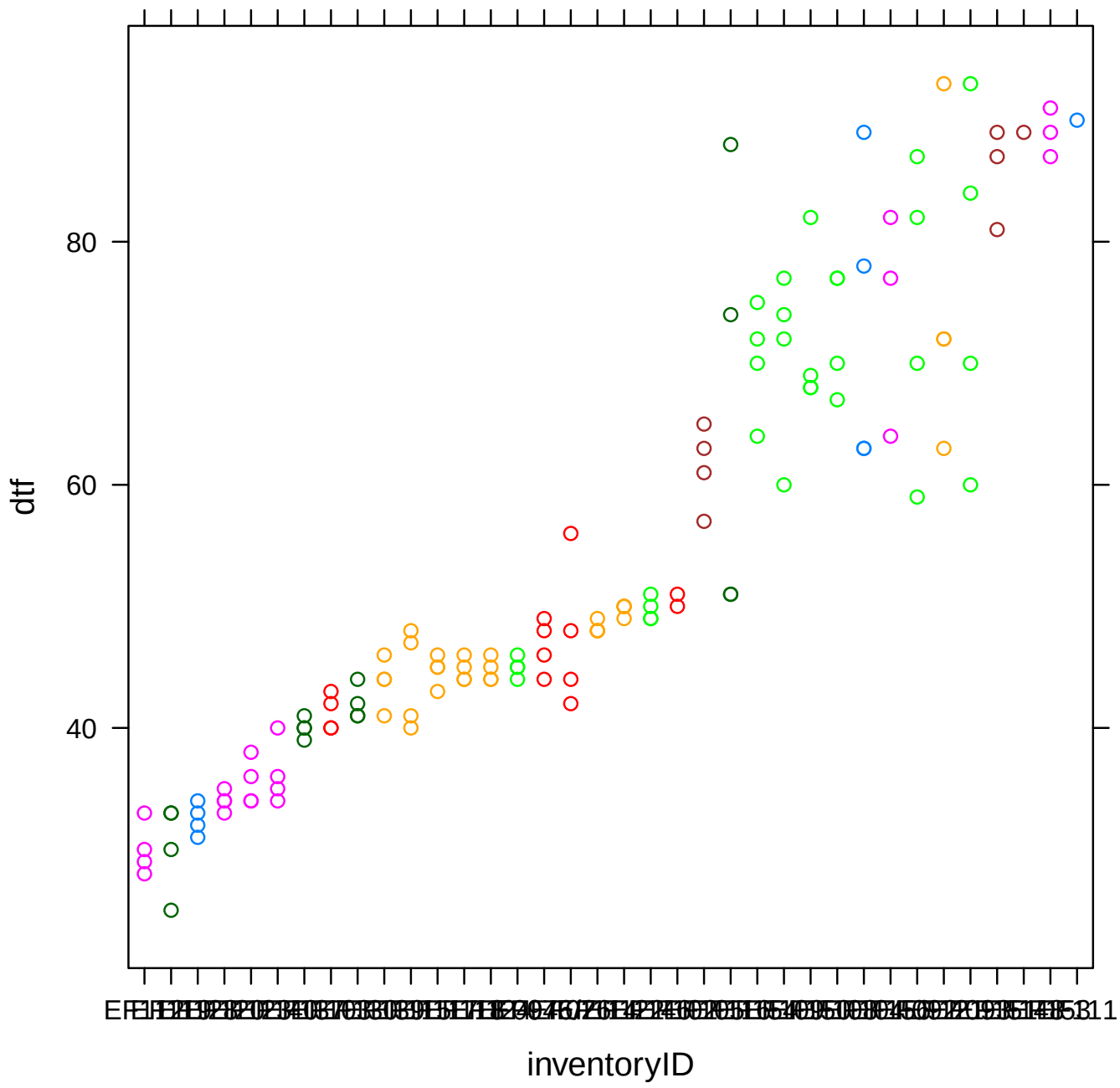
#--- dtf by packet, separately by subspecies ---------------
xyplot(logdtf~ inventoryID|subspecies, group=inventoryID,
                                       data = brassica2)
```

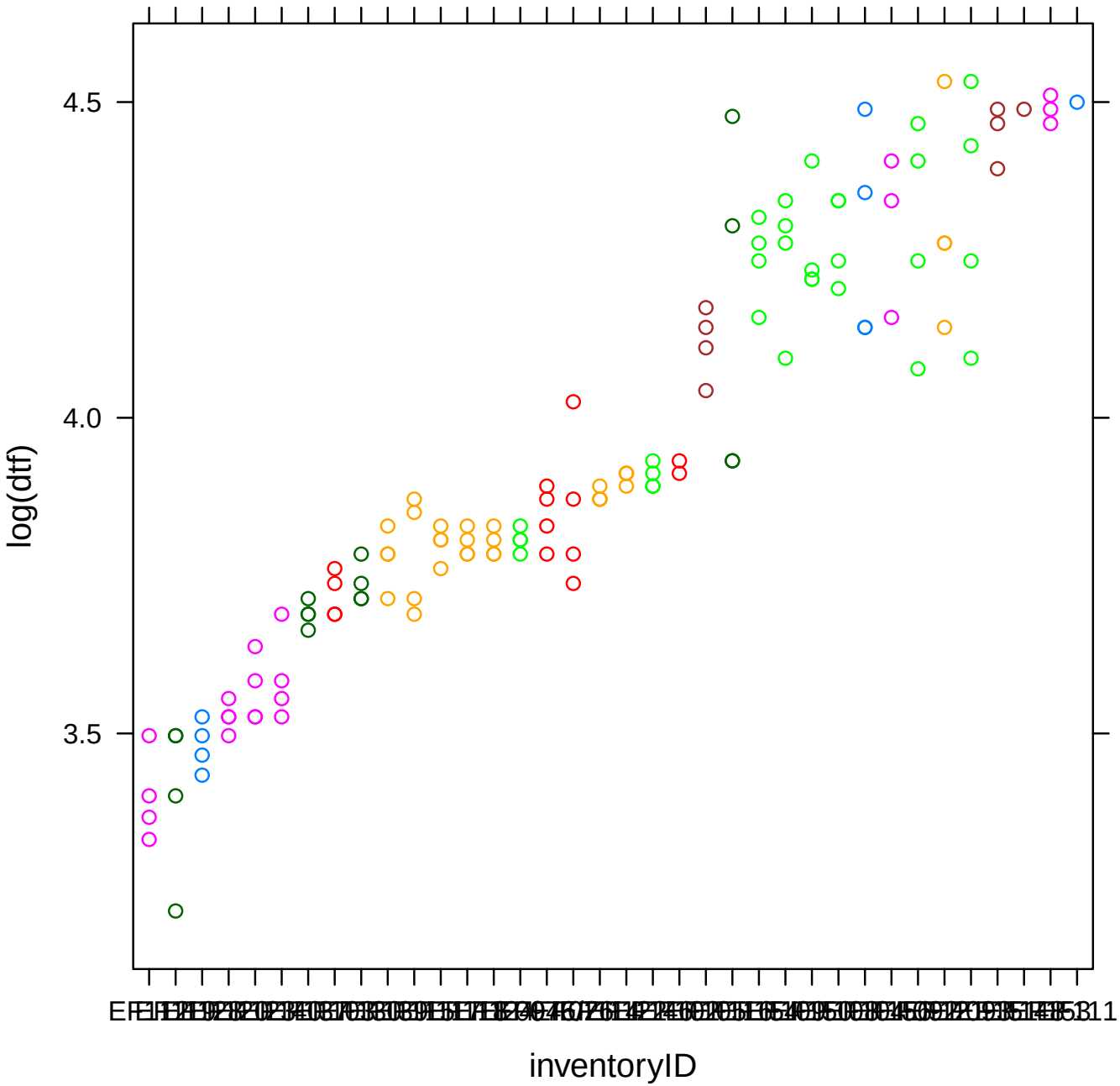

dtf
80
60
40
chinensis
dichotoma
marinosa
oleifera
parachinensis
pseudochinensis
spinosa
viridis
trilocularis
utilis
subspecies

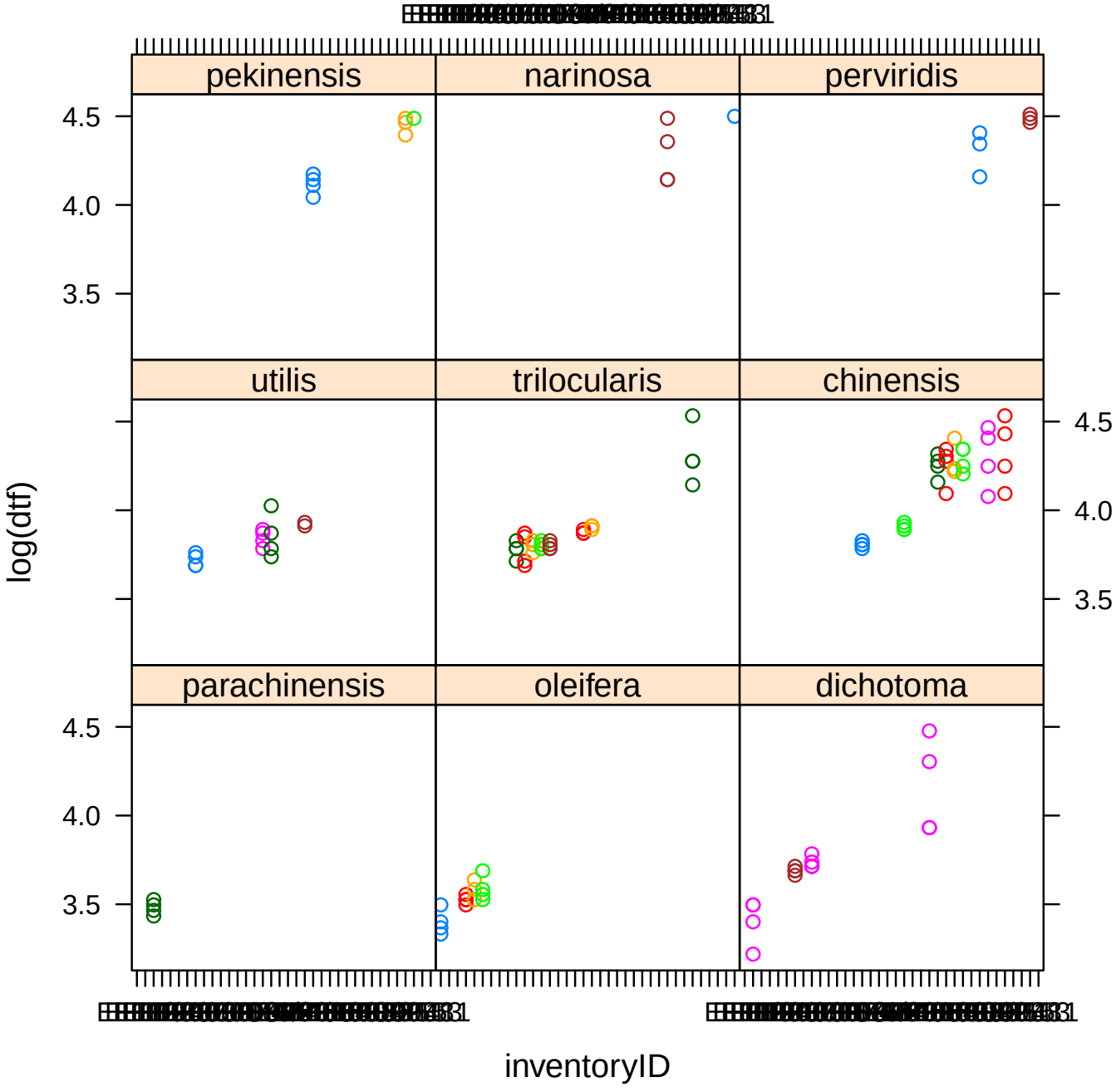

dtf
80
60
40
parachinensis
deifera
dichotoma
utilis
trilocularis
chinensis
pekinensis
narinosa
perviridis
subspecies


log(dtf)
4.5
4.0
3.5
parachinensis
deifera
dichotoma
utilis
trilocularis
chinensis
sekinensis
sarinosa
perviridis
subspecies


4.5
4.0
3.5
log(dtf)
inventoryID


dtf
80
60
40
inventoryID


4.5
4.0
3.5
log(dtf)
inventoryID


pekinensis
narinosa
perviridis
utilis
trilocularis
chinensis
parachinensis
oleifera
dichotoma
4.5
4.0
3.5
log(dtf)
inventoryID

Flowering time: Analysis with `lmer`

$$\log \text{dtf}_i = y_i = \beta_0 + \gamma_{k[j[i]]} + \alpha_{j[i]} + e_i$$

```
library(lme4)

fit=lmer(logdtf ~ (1|subspecies) + (1|inventoryID),
          data=brassica2)
```

`inventoryID`s were never repeated (`EF061, EF064`, etc.)
If instead `ID`s had been re-used across subspecies (ex: “1”, “2”, “3” for each subspecies), then we would need to use `subspecies:inventoryID` to indicate nesting:

```
fit=lmer(logdtf ~ (1|subspecies) + (1|subspecies:inventoryID),
          data=brassica2)
```

Flowering time: Results

```
> summary(fit)
Linear mixed model fit by REML
Formula: logdtf ~ (1 | subspecies) + (1 | inventoryID)
   Data: brassica2
    AIC    BIC logLik deviance REMLdev
 -115.5 -103.9  61.75   -126.0  -123.5

Random effects:
 Groups      Name        Variance  Std.Dev.
 inventoryID (Intercept) 0.0333514 0.182624
 subspecies  (Intercept) 0.1088915 0.329987
 Residual                0.0096426 0.098197
Number of obs: 133, groups: inventoryID, 36; subspecies, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)   3.9755     0.1163   34.17
```

Little residual variation within `inventoryID`.

Coefficients

```
> ranef(fit)  # random effects
$inventoryID  # alpha's
      (Intercept)
EF112   -0.145319
EF119   -0.337268
EF128   -0.114099
...
EF148    0.142729
EF153    0.139636
EF111    0.139726

$subspecies   # gamma's
             (Intercept)
parachinensis -0.372518
oleifera      -0.420390
dichotoma     -0.210588
utilis        -0.131054
trilocularis  -0.090316
chinensis      0.191628
pekinensis     0.329153
narinosa       0.344198
perviridis     0.359888

> fixef(fit) # fixed effects
(Intercept)
   3.975499
```

```
> coef(fit)  # estimated means:
$inventoryID # mu+gamma+alpha
        (Intercept)
EF112      3.830180
EF119      3.638231
EF128      3.861400
...
EF148      4.118229
EF153      4.115135
EF111      4.115226

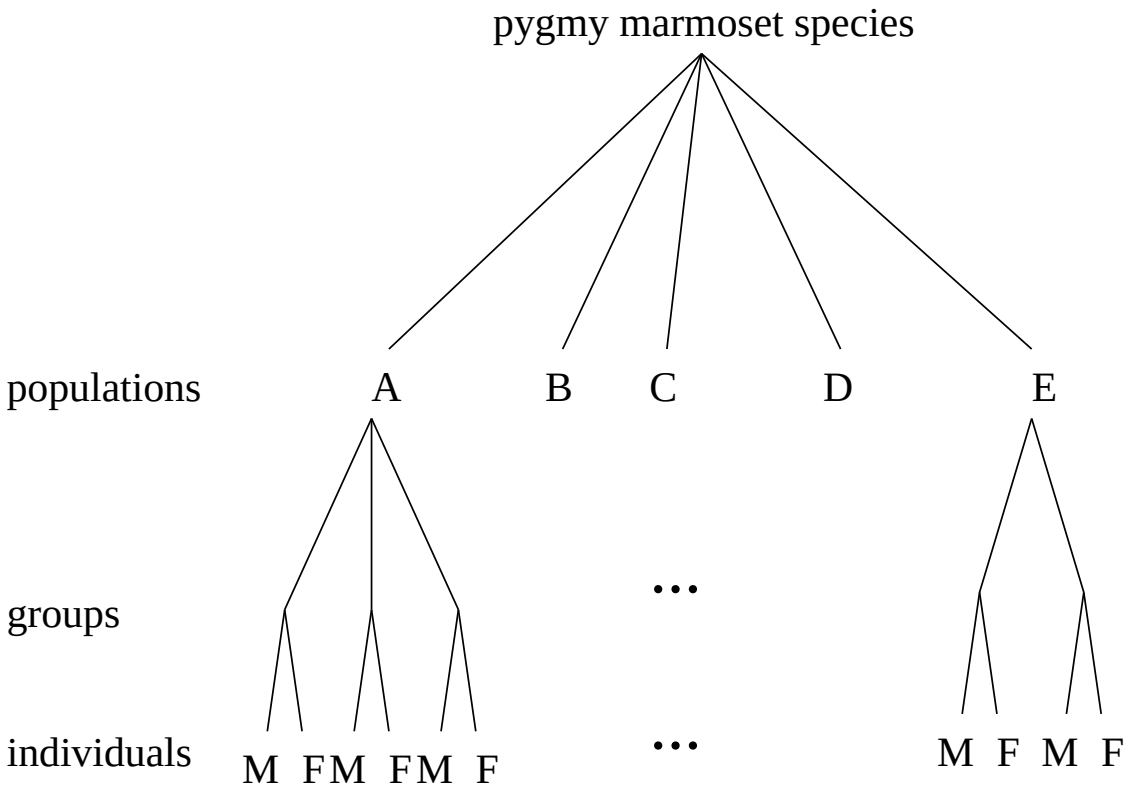
$subspecies  # mu + gamma
               (Intercept)
parachinensis    3.602981
oleifera         3.555109
dichotoma        3.764911
utilis           3.844444
trilocularis     3.885183
chinensis        4.167128
pekinensis       4.304653
narinosa         4.319697
perviridis       4.335388
```

Dialects in pygmy marmoset populations?

Vocal features were recorded and averaged over time for each indifidual. We will look at call duration (s) and note pitch (kHz) of “J calls”. S. de la Torre and C.T. Snowdon (2009)

- 5 populations were sampled from wild populations in northeastern Ecuador.
- 3 groups were sampled at random from each population. (2 groups in pop. E)
- One male and one female pygmy marmoset were sampled from each group.

Experimental design and potential predictors



Potential predictors:

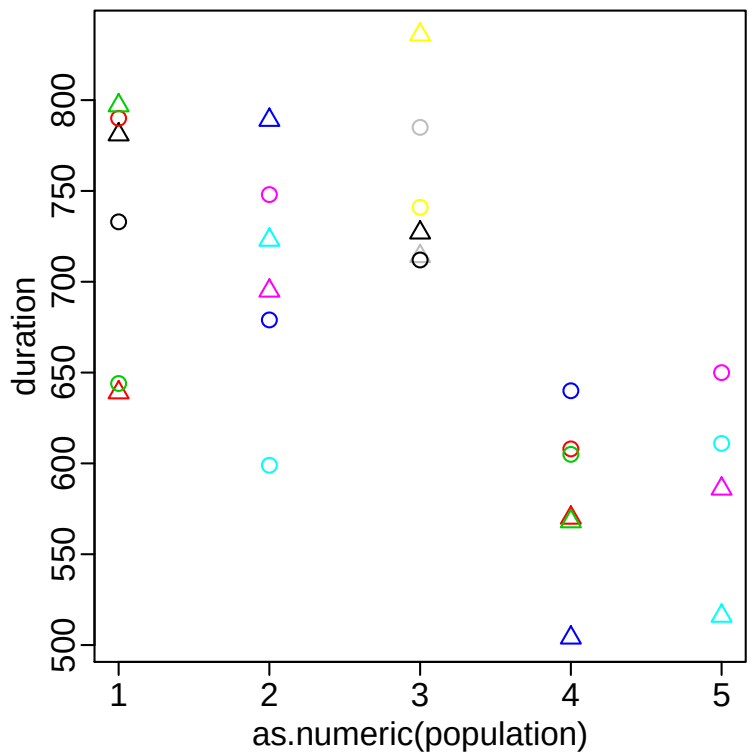
For which factors should we use fixed effects? random effects?

Fixed vs. Random effects for Group means

- Fixed group effects would take: 14-1 df. Out of a total of $n - 1 = 28 - 1$ df.
- Groups nested in populations: We cannot have both group effects and population effects unless group effects are random: we would want to estimate $14 - 1 = 13$ group effects and $5 - 1 = 4$ population effects. Too many parameters for the number of observed group means: 14.
- Fixed group effects would not constrain group means to be close to their population mean.
- Partial pooling will be very useful for estimating group effects: there are only 2 individuals per group (M and F). Not much info from each group. Estimated group means will preferably be close to their population mean.

Plot of call duration

```
> jcalls = read.table("marmosetJcalls.txt", header=T)
> str(jcalls)
'data.frame':   28 obs. of  8 variables:
 $ population: Factor w/ 5 levels "A","B","C","D",..: 3 3 3 3 3 3 4 4 4 4 ...
 $ group     : Factor w/ 14 levels "A1","A2","A3",..: 7 7 8 8 9 9 10 10 11 11 ...
 $ homerange : num  0.4 0.4 0.78 0.78 0.73 0.73 0.37 0.37 1.2 1.2 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 2 1 2 1 2 1 2 1 2 ...
 $ duration  : int  741 836 785 714 712 727 608 570 605 568 ...
 $ notes     : num  18.6 16.6 14.5 20 16.1 20.5 15.9 16.7 18.5 19.7 ...
 $ minFreq   : num  7.92 8.43 7.89 7.57 7.95 7.78 8.22 8.15 8.08 8.15 ...
 $ bandwidth : num  4.11 4.76 4.54 4.5 4.67 4.2 4.63 3.73 4.96 3.51 ...
> plot(duration ~ as.numeric(population), col=group, pch=as.numeric(sex), data=jcalls)
```



Mixed effect model to explain call duration

We will fit this model:

$$\text{call duration} = Y_i = \mu + \gamma_1 \mathbb{1}_{\text{male}} + \beta_{k[j[i]]} + \alpha_{j[i]} + e_i$$

where $i =$ individual #, $j =$ group #, $k =$ population ID, and

- γ_1 is fixed: represents the gender difference (gender effect)
- β_k are random, iid $\mathcal{N}(0, \sigma^2_\beta)$: represent population effects
- α_j are random, iid $\mathcal{N}(0, \sigma^2_\alpha)$: represent group effects
- e_i are random, iid $\mathcal{N}(0, \sigma^2_e)$: represent residual variability within groups, among individuals of the same group, same gender.

Mixed effect model to explain call duration

```
> library(lme4)
> fit.spg = lmer(duration ~ sex + (1|population) + (1|group),
                                                    data=jcalls)
> summary(fit.spg)
Linear mixed model fit by REML
Formula: duration ~ sex + (1 | population) + (1 | group)
   Data: jcalls
   AIC   BIC logLik deviance REMLdev
 311.2 317.9 -150.6    318.2   301.2
Random effects:
 Groups     Name        Variance Std.Dev.
 group      (Intercept)    0.0    0.000
 population (Intercept) 5717.3   75.613
 Residual                3624.9   60.207
Number of obs: 28, groups: group, 14; population, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  676.706     37.486  18.052
sexM          -7.143     22.756  -0.314

Correlation of Fixed Effects:
     (Intr)
sexM -0.304
```

Model with fixed population effects

Now consider:

$$\text{call duration} = Y_i = \mu + \gamma_1 \mathbb{1}_{\text{male}} + \beta_{k[j[i]]} + \alpha_{j[i]} + e_i$$

where $i =$ individual #, $j =$ group #, $k =$ population ID, and

- γ_1 is fixed: represents the gender difference (gender effect)
- β_k are fixed: still represent population effects (difference from a reference population for instance)
- α_j are random, iid $\mathcal{N}(0, \sigma_\alpha^2)$: represent group effects
- e_i are random, iid $\mathcal{N}(0, \sigma_e^2)$: represent residual variability within groups, among individuals of the same group, same gender.

Mixed with fixed population effects

```
> fit.spg = lmer(duration ~ sex + population + (1|group), data=jcalls)
> summary(fit.spg)
Linear mixed model fit by REML
Formula: duration ~ sex + population + (1 | group)
   Data: jcalls
   AIC   BIC logLik deviance REMLdev
 269.2 279.9 -126.6    302.2   253.2
Random effects:
 Groups   Name        Variance Std.Dev.
 group    (Intercept)    0.0    0.000
 Residual             3623.2   60.193
Number of obs: 28, groups: group, 14

Fixed effects:
            Estimate Std. Error t value
(Intercept)  734.238     27.079  27.115
sexM          -7.143     22.751  -0.314
populationB  -25.167     34.753  -0.724
populationC   21.833     34.753   0.628
populationD -148.167     34.753  -4.263
populationE -139.917     38.855  -3.601
```