Discussion7

Jing Lyu

2/22/2023

One-way ANOVA with random effects

We now consider situations where treatments are random samples from a large population of treatments. For example, we want to investigate the performance of randomly selected employee, or machines that were randomly sampled from a large population of machines. Then, we are interested in making a statement about some properties of the whole population and not of the observed individuals.

We use the following models to fit such data:

- Cell means model:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim_{iid} N(0, \sigma^2), \mu_i \sim_{iid} N(\mu, \sigma_\mu^2), \epsilon_{ij}, \mu_{i'}$ are independent for any $i, i' \in \{1, ..., r\}$

- Factor effects model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim_{iid} N(0, \sigma^2), \tau_i \sim_{iid} N(0, \sigma_\mu^2), \epsilon_{ij}, \tau_{i'}$ are independent for any $i, i' \in \{1, ..., r\}$

Properties:

$$E[Y_{ij}] = \mu, Var[Y_{ij}] = \sigma_\mu^2 + \sigma^2$$

$$Cor(Y_{ij}, Y_{kl}) = \begin{cases} 0 & \text{if } i \neq k \\ \sigma_\mu^2/(\sigma_\mu^2 + \sigma^2) & \text{if } i = k, j \neq l \\ 1 & \text{if } j = k, j = l \end{cases} \tag{1}$$

ICC Observations from different groups are uncorrelated while observations from the same group are correlated. We also call the correlation within the same group $\sigma_\mu^2/(\sigma_\mu^2 + \sigma^2)$ **intraclass correlation (ICC)**.

A large ICC indicates $\sigma_\mu^2 \gg \sigma^2$. Since σ^2 meansures the variation in the group and σ_μ^2 measures the variation between different groups, a large ICC means observations from the same group are much more similar than observations from different groups.

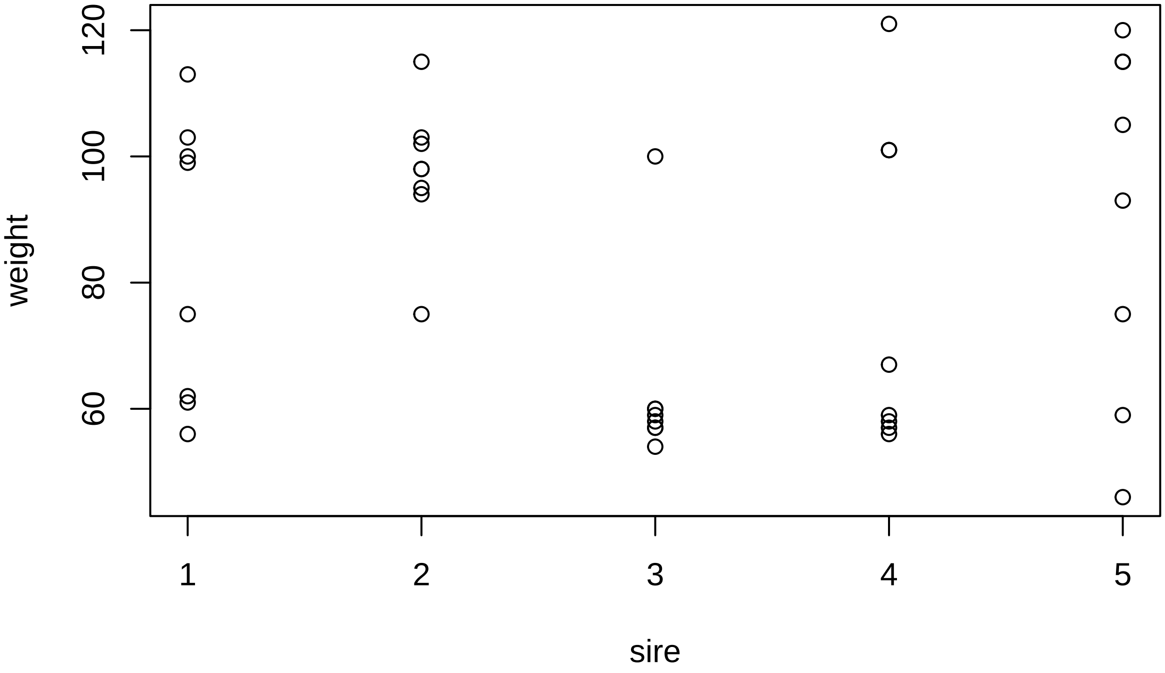
Now we consider an inheritance study (@kuehl2000designs) in which five sires, male animals, were each mated to a separate group of dams, female animals. The birth weights of eight male calves (from different dams) in each of the five sire groups were recorded. This is a balanced design with eight measurements for each sire.

```
## Create data set ####
weight = c(61, 100,  56, 113,  99, 103,  75,  62,  ## sire 1
           75, 102,  95, 103,  98, 115,  98,  94,   ## sire 2
           58,  60,  60,  57,  57,  59,  54, 100,   ## sire 3
           57,  56,  67,  59,  58, 121, 101, 101,   ## sire 4
           59,  46, 120, 115, 115,  93, 105,  75)   ## sire 5
```

```
sire = factor(rep(1:5, each = 8))
animals = data.frame(weight, sire)
str(animals)
```

```
## 'data.frame':    40 obs. of  2 variables:
##  $ weight: num  61 100 56 113 99 103 75 62 75 102 ...
##  $ sire  : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 2 ...
```

```
## Visualize data ####
stripchart(weight ~ sire, vertical = TRUE, pch = 1, xlab = "sire", data = animals)
```



Model fitting

- Model fitting with **aov**

We fitted the model as if it was a fixed effects model and then “adjusted” the output for random effects specific questions.

```
fit.aov = aov(weight ~ sire, data = animals)
summary(fit.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## sire         4   5591  1397.8   3.014 0.0309 *
## Residuals   35  16233   463.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows $MSTR = 1397.8, MSE = 463.8$. Then

$\hat{\sigma}^2_\mu = (MSTR - MSE)/n = (1397.8 - 463.8)/8 = 116.7.$

$\hat{\sigma}^2 = MSE = 463.8.$

- Model fitting with **lmer**

Now we want to use the more modern approach based on restricted maximum likelihood (REML) estimation technique.

```
library(lme4)
fit.animals = lmer(weight ~ (1 | sire), data = animals, REML = TRUE)
summary(fit.animals)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: weight ~ (1 | sire)
##    Data: animals
##
## REML criterion at convergence: 358.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9593 -0.7459 -0.1581  0.8143  1.9421
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  sire     (Intercept) 116.7    10.81
##  Residual             463.8    21.54
## Number of obs: 40, groups:  sire, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   82.550      5.911   13.96
```

Variance components are estimated by REML method. $\hat{\sigma}_\mu^2 = 116.7, \hat{\sigma}^2 = 463.8$.. ICC is $116.7/(116.7 + 463.8) \approx 0.2$ which indicates the variation between different sires is very small. You can say the proportion of variability that is due to variability in different sires is only 20%.

Since α_i are random quantities, not fixed parameters, we can get the conditional means (given the observed data) of α_i with `ranef`. Those conditional means are the best linear unbiased predictions, also known as **BLUPs**.

```
fixef(fit.animals) # extract the estimate of population mean: mu
```

```
## (Intercept)
##       82.55
```

```
sqrt(diag(vcov(fit.animals))) # standard error of the estimate of population mean
```

```
## (Intercept)
##    5.911403
```

```
ranef(fit.animals)
```

```
## $sire
##   (Intercept)
## 1   0.7183096
## 2   9.9895155
## 3 -12.9796882
## 4  -3.3743848
## 5   5.6462479
##
## with conditional variances for "sire"
```

```
VarCorr(fit.animals)
```

```
##  Groups   Name        Std.Dev.
##  sire     (Intercept) 10.805
##  Residual             21.536
```

Inference Main interest is to test whether random effects exist. The hypotheses are

$$H_0 : \sigma_\mu^2 = 0 \quad vs. \quad H_1 : \sigma_\mu^2 \neq 0$$

- F test

Test statistic is $MSTR/MSE \sim F(r-1,(n-1)r) = F(4,35)$. The observed value of test statistic is $MSTR/MSE = 3.0138$. P-value is $P(F_{4,35} > 3.0138) = 0.03087$.

- LR test

```
fit.animals = lmer(weight ~ (1 | sire), data = animals)
fit.animals.red = lm(weight ~ 1, data = animals)
anova(fit.animals, fit.animals.red)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: animals
## Models:
## fit.animals.red: weight ~ 1
## fit.animals: weight ~ (1 | sire)
##                 npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fit.animals.red    2 369.59 372.97 -182.79   365.59
## fit.animals        3 369.49 374.56 -181.75   363.49 2.0974  1     0.1475
```

Test statistic follows χ_1^2 distribution under the null. The observed value is 2.0974. P-value is $P(\chi_1^2 > 2.0974) = 0.1475$.

Suppose the significance level is 0.05, then we have different conclusions from the two tests.

Let's compare the 95% CIs for population mean under fixed and random effects model:

With `contr.sum`, we can obtain CI for population mean.

```
options(contrasts = c("contr.sum", "contr.poly"))
fit.animals.aov <- aov(weight ~ sire, data = animals)
confint(fit.animals.aov)["(Intercept)",] # CI for population mean from the fixed effects model
```

```
##    2.5 %   97.5 %
## 75.63725 89.46275
```

```
confint(fit.animals)["(Intercept)",] # CI for population mean from the random effects model
```

```
## Computing profile confidence intervals ...
```

```
##    2.5 %   97.5 %
## 69.83802 95.26197
```

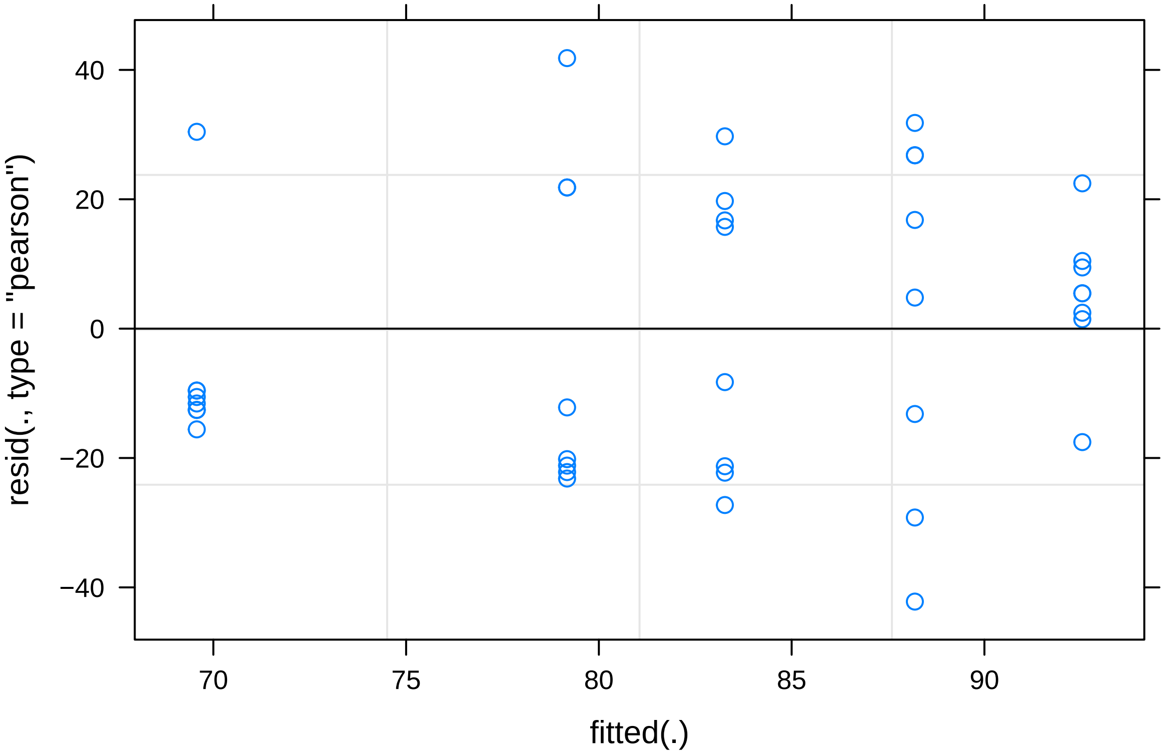
CI for population mean from the random effects model is **wider**.

One way of explaining this phenomenon might be that "The random effects model allows to make inference about the population of all sires (where we have seen five so far), while the fixed effects model allows to make inference about these five specific sires. Hence, we are facing a more difficult problem with the random effects model; this is why we are less confident in our estimate resulting in wider confidence intervals compared to the fixed effects model."

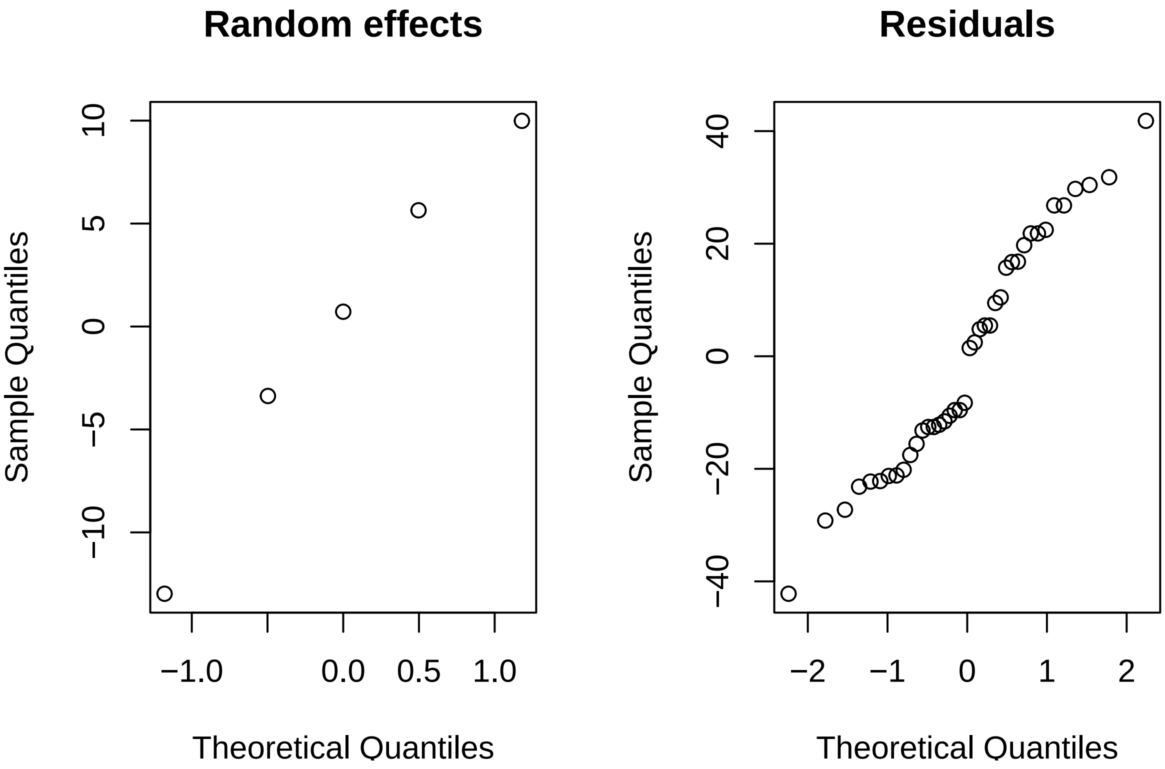
"From the theoretical perspective, these two tests should be asymptotically equivalent, and in final samples the F-test uses the exact distribution (if all assumptions hold…) while LRT is an approximation. In practice, however, it is unclear whether (1) sample size is large enough to invoke the asymptotic regime or (2) the normality assumption is reasonable enough to trust the F-test.” More investigation is needed if such discrepancy happens.

Model diagnostics We need to check the normality of the random effects and the residuals.

```
plot(fit.animals)
```



```
par(mfrow = c(1, 2))
qqnorm(ranef(fit.animals)$sire[,"(Intercept)"],
       main = "Random effects")
qqnorm(resid(fit.animals), main = "Residuals")
```



HW2 (9) Consider the following regression model

$$Y = X\beta + \epsilon, \tag{2}$$

where Y is a n_T-dimensional vector, X is an $n_T \times p$ matrix, β is a p-dimensional vector, and $\{\epsilon\} \sim$ MVN$(0, \sigma^2\mathrm{I})$, i.e., multivariate normal with covariance matrix $\sigma^2\mathrm{I}$. Express the residual sum of squares and explained sum of squares in Y and X, and then show that these two sum of squares are independent.

Hint:

$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$. We then have $\hat{Y} = X(X^T X)^{-1} X^T Y \equiv P_X Y$ and $e = Y - \hat{Y} = Y - X\hat{\beta} = (I - P_X)Y$. Therefore, we can write down the two sums of squares as

$$\mathrm{SSE} = e^T e = Y^T (I - P_X) Y \text{ and } \mathrm{SSR} = Y^T (P_X - P_1) Y,$$

where $P_1 = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T = \mathbf{1}\mathbf{1}^T/n_T, \mathbf{1} = (1, ..., 1)^T_{n_T}$.

$P_X, P_1, I - P_X, P_X - P_1$ are orthogonal projectors.

Properties of orthogonal projector: $P_X^T = P_X, P_X^T P_X = P_X, P_X^2 = P_X$.

Try to prove $cov((I - P_X)Y, (P_X - P_1)Y) = 0$. We will need $P_X P_1 = P_1$ where we use the fact that the first column of design matrix X is $(1, ..., 1)^T_{n_T}$. $\mathcal{C}(P_1) \in \mathcal{C}(X)$

References

https://stat.ethz.ch/~meier/teaching/anova/random-and-mixed-effects-models.html#eq:cell-means-random