Chapter 4 Analysis of Variance

4.3 Random and mixed effects

4.3.1 One-way ANOVA model with random effects

We now cosnider the following model in a balanced design, for $j = 1, \ldots, n, i = 1, \ldots, r$,

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

where ϵ_{ij} are i.i.d. $N(0, \sigma^2)$, μ_i are i.i.d. $N(\mu_\cdot, \sigma_\mu^2)$, and $\epsilon_{i,j}$ is independent with $\mu_{i'}$ for any $i, i' \in \{1, \ldots, r\}$. The total number of observations is $n_T = n \times r$.

The main change brought by the *random* effects is that the cell means are now *random variables*, instead of unobserved, fixed unknown parameters in the fixed effect model. As a result, it is no longer rigorous to say that we *estimate* the cell mean (we can't estimate a random variable!). Instead, we *predict* the cell means (as in predicting the *outcome* --- a random variable).

We have the following

$$\mathbb{E}[Y_{ij}] = \mu_\cdot,\ \mathrm{var}(Y_{ij}) = \mathrm{var}(\mu_i) + \mathrm{var}(\epsilon_{ij}) = \sigma_\mu^2 + \sigma^2,$$

and

$$\mathrm{cov}(Y_{ij}, Y_{ij'}) = \sigma_\mu^2,\ j \neq j', \mathrm{cov}(Y_{ij}, Y_{i'j'}) = 0,\ i \neq i'.$$

The factor-effect form is

$$Y_{ij} = \mu_\cdot + \tau_i + \epsilon_{ij},$$

where ϵ_{ij} are i.i.d. $N(0, \sigma^2)$, τ_i are i.i.d. $N(0, \sigma_\tau^2)$, and ϵ_i is independent with $\tau_{i'}$ for any $i, i' \in \{1, \ldots, r\}$. Here $\tau_i = \mu_i - \mu_\cdot$ and $\sigma_\tau = \sigma_\mu$.

4.3.2 Sum of squares

We can decompose the sum of squres as

$$\mathrm{SSTO} = \sum_{i=1}^{r} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2,\ df = nr - 1,$$

$$\mathrm{SSTR} = n \sum_{j=1}^{r} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2,\ df = r - 1,\ \mathrm{MSTR} = \frac{\mathrm{SSTR}}{r-1}.$$

$$\mathrm{SSE} = \sum_{i=1}^{r} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2,\ df = r(n-1),\ \mathrm{MSE} = \frac{\mathrm{SSE}}{r(n-1)}.$$

Properties

1. $\mathbb{E}\left[\bar{Y}_{\cdot\cdot}\right] = \mu_\cdot$. We can estimate the overall (population) mean $\mu_\cdot$ with $\hat{\mu}_\cdot = \bar{Y}_{\cdot\cdot}$.
2. $\mathrm{var}\left(\bar{Y}_{\cdot\cdot}\right) = (n\sigma_\mu^2 + \sigma^2)/(nr)$. We have $\hat{\mathrm{var}}(\bar{Y}) = \mathrm{MSTR}/(nr)$.

3. $\mathbb{E}[\mathrm{MSE}] = \sigma^2$. We can estimate σ^2 with $s^2 = \mathrm{MSE}$.

4. $\mathbb{E}[\mathrm{MSTR}] = n\sigma_\mu^2 + \sigma^2$. We can estimate σ_μ^2 with $s_\mu^2 = (\mathrm{MSTR} - \mathrm{MSE})/n$,

Best linear unbiased *predictor* of τ_i

Consider a linear predictor of τ_i of the form $H = d_1\bar{Y}_{1\cdot} + \cdots + d_r\bar{Y}_{r\cdot}$ s.t. $\mathbb{E}[H - \tau_i] = 0$. Then, $\mathbb{E}\left[(H - \tau_i)^2\right]$ is minimized when $d_i = w(1 - 1/r)$ and $d_{i'} = -w/r$ for $i' \neq i$. The best predictor is of the form $w(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})$ with $w = n\sigma_\mu^2/(n\sigma_\mu^2 + \sigma^2)$, and thus $\hat{\tau}_o = \hat{w}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})$ with $\hat{w} = ns_\mu^2/\mathrm{MSTR}$.

Remarks:

1. $\hat{w} \in [0, 1]$.
2. $\mathrm{SSE} = \sum\sum(Y_{ij} - \bar{Y}_{i\cdot})^2$ is free of μ and τ's.
3. $H = \sum\sum c_{ij}Y_{ij}$.

4.3.3 Inference with random effects

From the factor effect form, we can see that the expectation of random effects can be absorbed in the overall mean. Hence, when random effects do not exist, it means that the variance σ_μ^2 is zero. We have the following hypotheses

$$H_0 : \sigma_\mu^2 = 0 \text{ v.s. } H_1 : \sigma_\mu^2 \neq 0.$$

The test statistic remains the same as $F^* = \mathrm{MSTR}/\mathrm{MSE}$, which follows an F-distribution with d.f. $(r - 1, (n - 1)r)$ under the null.

Under the null hypothesis, the ratio $\sigma_\mu^2/(\sigma_\mu^2 + \sigma^2)$ equals zero. Otherwise, the ratio $\sigma_\mu^2/(\sigma_\mu^2 + \sigma^2)$ takes values in $(0, 1]$. This ratio characterizes the amount of variability in the samples that can be explained by the random effects from the factor. We can estimate this ratio with $s_\mu^2/(s_\mu^2 + \mathrm{MSE})$.

However, the construction of its confidence interval is slightly more involved. The confidence interval for this ratio is $(L/(L+1), U/(U+1))$, where

$$L = \frac{1}{n}\left[\frac{F^*}{F(1 - \alpha/2; r - 1, (n-1)r)} - 1\right], U = \frac{1}{n}\left[\frac{F^*}{F(\alpha/2; r - 1, (n-1)r)} - 1\right],$$

and $F^* = \mathrm{MSTR}/\mathrm{MSE}$. For σ_μ^2/σ^2, the confidence interval is (L, U).

In addition, we have the following confidence intervals.

For σ^2, the conferences interval is (L, U) where

$$L = \frac{df\mathrm{MSE}}{\chi^2(1 - \alpha/2; df)}, U = \frac{df\mathrm{MSE}}{\chi^2(\alpha/2; df)}.$$

For σ_μ^2, we can use the Scatterthwaite approximation to the distribution of s_μ^2. To this end, we view σ_μ^2 as

$$\sigma_\mu^2 = \frac{1}{n}\mathbb{E}[\mathrm{MSTR}] - \frac{1}{n}\mathbb{E}[\mathrm{MSE}] \equiv c_1\mathbb{E}[\mathrm{MSTR}] + c_2\mathbb{E}[\mathrm{MSE}].$$

Let $T = c_1\mathbb{E}[\mathrm{MS}_1] + \cdots + c_j\mathbb{E}[\mathrm{MS}_h]$, $\hat{T} = c_1\mathrm{MS}_1 + \cdots + c_j\mathrm{MS}_h$, and

$$df = \frac{\hat{T}^2}{(c_1\mathrm{MS}_1)^2/df_1 + \cdots + (c_h\mathrm{MS}_h)^2/df_h}.$$

We have $(df)\hat{T}/T \approx \chi^2_{df}$. We can construct the confidence interval accordingly.

4.3.4 Two-way ANOVA with random effects

Consider a two-way ANOVA model with balanced design and random effects, for $k = 1, \ldots, n, j = 1, \ldots, b, i = 1, \ldots, a,$

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where (i) $\alpha_i \sim N(0, \sigma^2_\alpha)$, (ii) $\beta_j \sim N(0, \sigma^2_\beta)$, (iii) $(\alpha\beta)_{ij} \sim N(0, \sigma^2_{\alpha\beta})$, (iv)$\{\epsilon_{ijk}\} \sim N(0, \sigma^2)$, and (v) all random variables are mutually independent.

In this model, we have the following results, similar to the case with one factor.

1. $\mathbb{E}[Y_{ijk}] = \mu_{..}$.
2. $\mathrm{var}(Y_{ijk}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta} + \sigma^2$.
3. $\mathrm{cov}(Y_{ijk}, Y_{ijk'}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta}$ for $k \neq k'$.
4. $\mathrm{cov}(Y_{ijk}, Y_{ij'k'}) = \sigma^2_\alpha$ for $j \neq j'$.
5. $\mathrm{cov}(Y_{ijk}, Y_{i'jk'}) = \sigma^2_\beta$ for $i \neq i'$.
6. $\mathrm{cov}(Y_{ijk}, Y_{i'j'k'}) = 0$ when no indices are equal.

Estimation of the variance components. We have $\mathrm{SSA} = nb\sum\hat{\alpha}_i^2, \mathrm{SSB} = na\sum\hat{\beta}_j^2, \mathrm{SSAB} = n\sum\sum\hat{(\alpha\beta)}_{ij}^2$. We can see that $\mathbb{E}[\mathrm{MSAB}] - \mathbb{E}[\mathrm{MSE}] = n\sigma^2_{\alpha\beta}, \mathbb{E}[\mathrm{MSA}] - \mathbb{E}[\mathrm{MSAB}] = nb\sigma^2_\alpha, \mathbb{E}[\mathrm{MSB}] - \mathbb{E}[\mathrm{MSAB}] = na\sigma^2_\beta$. Hence, we can define unbiased estimators of the variance components based on the mean squares. Furthermore, we have $\mathbb{E}[\bar{Y}] = \mu$, $\mathrm{var}(\bar{Y}) = (nb\sigma^2_\alpha + na\sigma^2_\beta + n\sigma^2_{\alpha\beta} + \sigma^2)/(nab)$.

The analytic solutions for the unbalanced design are beyond the scope of this note. We can still fit the model in its regression form with the *restricted maximum likelihood*. Theory, properties, and testing follow accordingly, see, e.g., this note.

4.3.5 Mixed effects

When fixed effects and random effects are both present, our model takes the same form but there are a few more assumptions

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ k = 1, \ldots, n, j = 1, \ldots, b, i = 1, \ldots, a,$$

where (i) $\sum\alpha_i = 0$, (ii) β_j are i.i.d. $N(0, \sigma^2_\beta)$, (iii) $\sum_i(\alpha\beta)_{ij} = 0$ for any j, (iv) $(\alpha\beta)_{ij} \sim N(0, (1 - 1/a)\sigma^2_{\alpha\beta})$, (v) $\mathrm{cov}((\alpha\beta)_{ij}, (\alpha\beta)_{i'j}) = -\sigma^2_{\alpha\beta}/a$, (vi) $\mathrm{cov}((\alpha\beta)_{ij}, (\alpha\beta)_{i'j'}) = 0$, if $i \neq i'$ and $j \neq j'$, (vii)$\{\epsilon_{ijk}\}$ are i.i.d. $N(0, \sigma^2)$, and (viii) $\{\beta_j\}$, $\{(\alpha\beta)_{ij}\}$, $\{\epsilon_{ijk}\}$ are mutually independent.

In this model, we have

1. $\mathbb{E}[Y_{ijk}] = \mu_{..} + \alpha_i$

2. $\mathrm{var}(Y_{ijk}) = \sigma_\beta^2 + (1 - 1/a)\sigma_{\alpha\beta}^2 + \sigma^2$.
3. $\mathrm{cov}(Y_{ijk}, Y_{ijk'}) = \sigma_\beta^2 + (1 - 1/a)\sigma_{\alpha\beta}^2$ for $k \neq k'$.
4. $\mathrm{cov}(Y_{ijk}, Y_{ij'k'}) = 0$ for $j \neq j'$.
5. $\mathrm{cov}(Y_{ijk}, Y_{i'jk'}) = \sigma_\beta^2 - \sigma_{\alpha\beta}^2/a$ for $i \neq i'$.
6. $\mathrm{cov}(Y_{ijk}, Y_{i'j'k'}) = 0$ when no indices are equal.

We can see that $\mathbb{E}[\mathrm{MSAB}] - \mathbb{E}[\mathrm{MSE}] = n\sigma_{\alpha\beta}^2, \mathbb{E}[\mathrm{MSB}] - \mathbb{E}[\mathrm{MSAB}] = na\sigma_\beta^2$. We can define unbiased estimators of the variance components based on the mean squares.

We have $\hat{H} = \sum c_i \bar{Y}_{i\cdot\cdot}$, $\mathrm{var}(\hat{H}) = \sum c_i^2 (n\sigma_{\alpha\beta}^2 + \sigma^2)/(nb)$, and $s^2(\hat{H}) = \sum c_i^2 \mathrm{MSAB}/(nb)$.

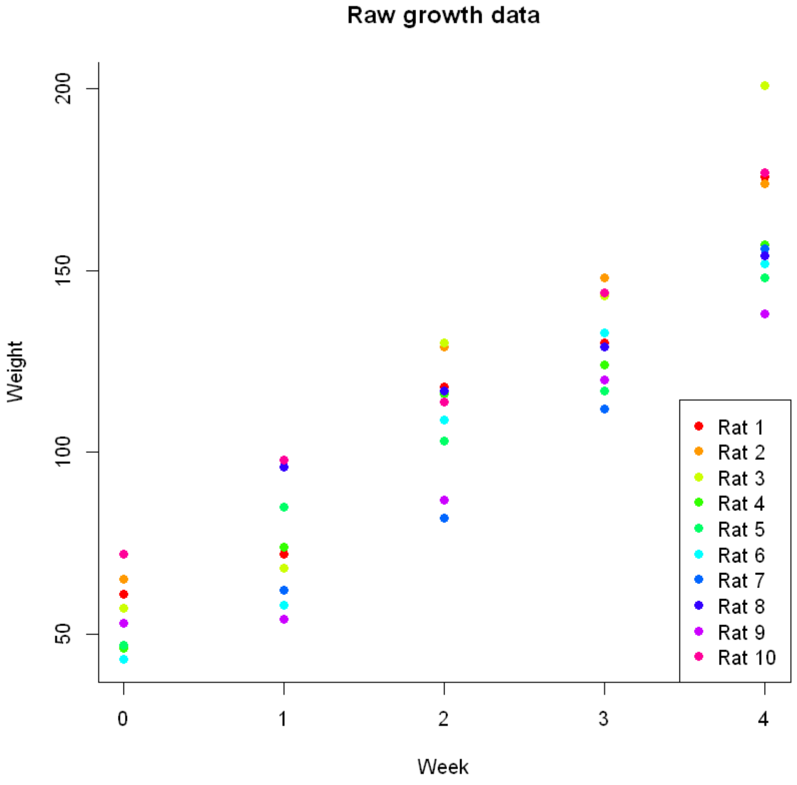
4.3.6 Example: rat growth data

We consider the rat growth data. Each rat is measured over 5 weeks. This type of data set is called longitudinal since the observations are taken over time. There is a covariate `mother's weight` (X). We consider several models to fit them in `R`. For more on the syntax of `lmer` see the vignette here.

In [2]:

```
Rat.growth <- read.csv(file="../Data/Growth.csv", header=TRUE, sep=",")

colorpicks = rainbow(n=length(unique(Rat.growth$rat)));
with(Rat.growth, plot(y=weight, x=week,type='p', pch=16, bty='l',  main='Raw growth d
for(i in 1:length(unique(Rat.growth$rat))){
  one.rat=Rat.growth[Rat.growth$rat==i,]
  with(one.rat, lines(weight, week,col=colorpicks[rat]))
}
# For more thoughts on visualization, see http://www.colbyimaging.com/wiki/statistics/
legend('bottomright', col=colorpicks, pch=c(16), legend=paste('Rat', unique(Rat.growt
```



In [5]:

```
library(lme4)
lm1=lmer(weight~as.factor(week)+(1|rat),data=Rat.growth)
lm2=lmer(weight~week+X+(1|rat),data=Rat.growth)
lm3=lmer(weight~as.factor(week)+X+(1|rat),data=Rat.growth)
lm4=lmer(weight~week+X+(1|rat)+(0+week|rat),data=Rat.growth)
lm5=lmer(weight~week+I(week^2)+X+(1|rat),data=Rat.growth)
lm6=lmer(weight~week+I(week^2)+X+(1|rat)+(0+week|rat)+(0+week^2|rat),data=Rat.growth

# For model selection, we can use AIC, BIC
AIC(lm1,lm2,lm3)

BIC(lm1,lm2,lm3,lm4,lm5,lm6)
```

```
Warning message in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
"unable to evaluate scaled gradient"
Warning message in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
"Model failed to converge: degenerate  Hessian with 1 negative eigenvalues"
```

A data.frame: 3 × 2

	df	AIC
	<dbl>	<dbl>
lm1	7	379.5793
lm2	5	390.3823
lm3	8	373.3366

A data.frame: 6 × 2

	df	BIC
	<dbl>	<dbl>
lm1	7	392.9635
lm2	5	399.9424
lm3	8	388.6328
lm4	6	403.8459
lm5	6	400.9624
lm6	8	408.7746