

Random Effects Models

If the levels of a factor are chosen at **random** from a larger population of levels, then the factor is said to be a random factor. Thus we would like to incorporate this information into the model via a random effects approach. As we saw in chapter 5, random effects can also be used to model correlation structures in the data. Note that the 4 fields were randomly chosen from a larger population of such fields, so random effects based on that criteria is appropriate.

- **One-Factor Random Effects Model:**

Example (Textile — from Montgomery): “A textile mill has a large number of looms. Each loom is supposed to provide the same output of cloth per minute. To investigate the assumption, five looms are chosen at random, and their output is noted at different times.”

loom 1	14.0	14.1	14.2	14.0	14.1
loom 2	13.9	18.8	13.9	14.0	14.0
loom 3	14.1	14.2	14.1	14.0	13.9
loom 4	13.6	13.8	14.0	13.9	13.7
loom 5	13.8	13.6	13.9	13.8	14.0

Experimental material: Yarn used to create the cloth.

Experimental design:

- The yarn is randomly distributed to the looms.
- Then 5 looms are randomly sampled to take part in the experiment.
- For each loom, 5 repetitions of output at varying times was noted. Then from those 5 replications, 5 output (lb/minute) readings were obtained.

Consider the following model:

$$y_{ij} = \mu + \tau_i + \epsilon_{i,j}; \quad \tau_i \stackrel{\text{iid}}{\sim} \text{normal}(0, \text{variance} = \sigma_\tau^2), \quad \epsilon_{ij} \stackrel{\text{iid}}{\sim} \text{normal}(0, \text{variance} = \sigma^2)$$
$$i = 1, \dots, t_a = 5; \quad j = 1, \dots, r = 5.$$

Here, τ_i is the random effect for loom i . Note the following:

$$E(y_{ij}) = \mu$$
$$V(y_{ij}) = \sigma_\tau^2 + \sigma^2$$

With the model we also create correlation of data from the same loom — as was the case with the 4 fields in chapter 5:

$$\begin{aligned} \text{Cov}(y_{ij_1}, y_{ij_2}) &= E[(y_{ij_1} - E[y_{ij_2}]) \times (y_{ij_2} - E[y_{ij_2}])] \\ &= E[(\mu + \tau_i + \epsilon_{i,j_1} - \mu) \times (\mu + \tau_i + \epsilon_{i,j_2} - \mu)] \\ &= E[\tau_i^2 + \tau_i \times (\epsilon_{i,j_1} + \epsilon_{i,j_2}) + \epsilon_{i,j_1} \times \epsilon_{i,j_2}] \\ &= E[\tau_i^2] + 0 + 0 = \sigma_\tau^2 \end{aligned}$$

Note that there is no correlation across looms!

$$\begin{aligned}
 Cov(y_{i_1j}, y_{i_2j}) &= E[(y_{i_1j} - E[y_{i_1j}]) \times (y_{i_2j} - E[y_{i_2j}])] \\
 &= E[(\mu + \tau_{i_1} + \epsilon_{i_1,j} - \mu) \times (\mu + \tau_{i_2} + \epsilon_{i_2,j} - \mu)] \\
 &= E[\tau_{i_1}\tau_{i_2} + \tau_{i_1}\epsilon_{i_1,j} + \tau_{i_2}\epsilon_{i_1,j} + \epsilon_{i_1,j}\epsilon_{i_2,j}] \\
 &= 0 + 0 + 0 + 0 = 0
 \end{aligned}$$

To determine whether at least one the *looms* is different from the others, we now focus our hypothesis test on the variance term σ_τ^2 :

$$\begin{aligned}
 H_0 : \sigma_\tau^2 &= 0 \\
 H_1 : \sigma_\tau^2 &> 0
 \end{aligned}$$

We can see that if $\sigma_\tau^2 = 0$ then all the looms are the same, and if $\sigma_\tau^2 > 0$ there is variability among the looms. Now we can consider the ANOVA decomposition — which still holds: $SS_{total} = SS_{treatment} + SS_{error}$:

$$\begin{aligned}
 SS_{total} &= \sum_{i=1}^{t_a} \sum_{j=1}^r (y_{ij} - \bar{y}_{..})^2 \\
 SS_{treatment} &= \sum_{i=1}^{t_a} \sum_{j=1}^r (\bar{y}_{i.} - \bar{y}_{..})^2 \\
 SS_{error} &= \sum_{i=1}^{t_a} \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2
 \end{aligned}$$

Let's determine the $E[MS_{treatment} = SS_{treatment}/(t_a - 1)]$. Let's start with the $E[SS_{treatment}]$:

$$\begin{aligned}
E[SS_{treatment}] &= E \left[\sum_{i=1}^{t_a} \sum_{j=1}^r (\bar{y}_{i.} - \bar{y}_{..})^2 \right] \\
&= E \left[\sum_{i=1}^{t_a} \sum_{j=1}^r (\bar{y}_{i.}^2 - 2\bar{y}_{..}\bar{y}_{i.} + \bar{y}_{..}^2) \right] \\
&= E \left[\sum_{i=1}^{t_a} r\bar{y}_{i.}^2 - 2r\bar{y}_{..} \sum_{i=1}^{t_a} \bar{y}_{i.} + t_a r\bar{y}_{..}^2 \right] \\
&= E \left[\sum_{i=1}^{t_a} r\bar{y}_{i.}^2 - 2rt_a\bar{y}_{..} \sum_{i=1}^{t_a} \frac{\bar{y}_{i.}}{t_a} + t_a r\bar{y}_{..}^2 \right] \\
&= E \left[\sum_{i=1}^{t_a} r\bar{y}_{i.}^2 - 2rt_a\bar{y}_{..}\bar{y}_{..} + t_a r\bar{y}_{..}^2 \right] \\
&= E \left[\sum_{i=1}^{t_a} r\bar{y}_{i.}^2 - 2rt_a\bar{y}_{..}^2 + t_a r\bar{y}_{..}^2 \right] \\
&= E \left[\sum_{i=1}^{t_a} r\bar{y}_{i.}^2 - rt_a\bar{y}_{..}^2 \right] \\
&= rE \left[\sum_{i=1}^{t_a} \bar{y}_{i.}^2 \right] - rt_a E [\bar{y}_{..}^2]
\end{aligned}$$

So let's figure out:

$$E[\bar{y}_{i.}] = ?$$

$$V[\bar{y}_{i.}] = ?$$

$$E[\bar{y}_{i.}^2] = ?$$

$$E[\bar{y}_{..}] = ?$$

$$V[\bar{y}_{..}] = ?$$

$$E[\bar{y}_{..}^2] = ?$$

Now:

$$E[SS_{treatment}] = ?$$

$$E[MS_{treatment}] = ?$$

So under the null hypothesis H_0 , The $E(MS_{treatment}) = \sigma^2$. Also, as before, the $E(MS_{error}) = \sigma^2$ both under H_0 and H_1 — thus the MS_{error} is an unbiased estimator of σ^2 . Also, as before:

$$\begin{aligned} SS_{treatment}/\sigma^2 &\sim \chi_{t_a-1}^2 \\ SS_{error}/\sigma^2 &\sim \chi_{t_a r - t_a}^2 \\ MS_{treatment}/MS_{error} &\sim F_{t_a-1, t_a r - t_a} \end{aligned}$$

So, the ANOVA table, including the p-values, is determined the same way! However, the conclusions apply to the entire population, since we took a random sample of looms! So we have:

```
> fit.lm <- lm(y~ as.factor(loom))
> anova(fit.lm)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
as.factor(loom)  4  4.1816   1.0454   1.0949  0.3860
Residuals      20 19.0960   0.9548
```

Now, we would like to also get estimates for σ^2 and σ_τ^2 . We will use the *method of moments* to estimate the parameters. We will set the empirical moment equal to the theoretical moment based on the model. In this case we have two parameters, so we will need two equations and we will base them on the empirical and theoretical means:

$$\begin{aligned} MS_{error} &= \sigma^2 \\ MS_{treatment} &= \sigma^2 + r\sigma_\tau^2 \end{aligned}$$

So:

$$\begin{aligned} \hat{\sigma}^2 &= MS_{error} \\ \hat{\sigma}_\tau^2 &= (MS_{treatment} - MS_{error})/r \end{aligned}$$

```
> sigma.sq.hat <- anova(fit.lm)[2,3]
> sigma.sq.tau.hat <- (anova(fit.lm)[1,3] - anova(fit.lm)[2,3])/5
>
> sigma.sq.hat
[1] 0.9548
> sigma.sq.tau.hat
[1] 0.01812
```

We can also use the *nlme* package in R, that was used in chapter 5. This estimates the parameters of the model using *restricted maximum likelihood (REML) estimation*. We will not go through that procedure in this course:

```
> library(nlme)
> fit.me <- lme(fixed=y~ +1, random=~1|as.factor(loom))
> fit.me
```

```

Linear mixed-effects model fit by REML
Data: NULL
Log-restricted-likelihood: -35.29023
Fixed: y ~ +1
(Intercept)
    14.136

```

```

Random effects:
Formula: ~1 | as.factor(loom)
(Intercept) Residual
StdDev:    0.1346108 0.9771387

```

```

Number of Observations: 25
Number of Groups: 5

```

So our REML estimates are: $\hat{\sigma}_{reml}^2 = 0.977^2 = 0.955$ and $\hat{\sigma}_{\tau_{reml}}^2 = 0.135^2 = 0.018$ So we see that most of the variation is not due to the differences between looms, and specifically we can't say that $\sigma_{\tau} > 0$, from the hypothesis test!

- **Two-Factor Mixed Effects Model:**

Example (Carbon Anode — from Montgomery): “An experiment was conducted to determine whether either firing temperature or furnace position affect the baked density of a carbon anode.” Assume that the two positions were randomly selected, but the temperatures are fixed.

Experimental material: the carbon used in the experiment.

Experimental design:

- 2 positions are randomly sampled.
- The carbon is randomly distributed to the position \times temperature cells.
- There are 3 replicates for each cell.

	temp (800)	temp (825)	temp (850)
pos 1	570	1063	565
	565	1080	510
	583	1043	590
pos 2	528	988	526
	547	1026	538
	521	1004	532

Let's consider the following model:

$$\begin{aligned}
 y_{ijk} &= \mu + a_i + b_j + (ab)_{ij} + \epsilon_{ijk}, \\
 i &= \{1, \dots, t_a = 3\}; j = \{1, \dots, t_b = 2\}; k = \{1, \dots, r = 3\}
 \end{aligned}$$

Where $b_j \stackrel{\text{iid}}{\sim} \text{normal}(0, \text{variance} = \sigma_b^2)$ and the interaction $(ab)_{ij} \stackrel{\text{iid}}{\sim} \text{normal}(0, \text{variance} = [(t_a - 1)/t_a] \sigma_{ab}^2)$. The first thing to note is that since the interaction consists of a random variable it is random. Next, the factor $[(t_a - 1)/t_a]$ on the variance is used to **simplify** the expected mean squares. What else is needed to specify the model:

- $\sum_{i=1}^{t_a} a_i = 0$
- $\sum_{i=1}^{t_a} (ab)_{ij} = (ab)_{.j} = 0$

We have the following sum of squares decompositions:

$$\begin{aligned}
\sum_{i=1}^{t_a} \sum_{j=1}^{t_b} \sum_{k=1}^r (y_{ijk} - \bar{y}_{...})^2 &= \sum_{i=1}^{t_a} \sum_{j=1}^{t_b} \sum_{k=1}^r (\bar{y}_{i..} - \bar{y}_{...})^2 \\
&+ \sum_{i=1}^{t_a} \sum_{j=1}^{t_b} \sum_{k=1}^r (\bar{y}_{.j.} - \bar{y}_{...})^2 \\
&+ \sum_{i=1}^{t_a} \sum_{j=1}^{t_b} \sum_{k=1}^r (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\
&+ \sum_{i=1}^{t_a} \sum_{j=1}^{t_b} \sum_{k=1}^r (\bar{y}_{ijk} - \bar{y}_{ij.})^2
\end{aligned}$$

Now we have the following expected mean squares:

- $E(MSA) = \sigma^2 + r\sigma_{ab}^2 + \frac{t_b r \sum_{i=1}^{t_a} a_i^2}{t_a - 1}$
- $E(MSB) = \sigma^2 + t_a r \sigma_b^2$
- $E(MSAB) = \sigma^2 + r\sigma_{ab}^2$
- $E(MSE) = \sigma^2$

Based on this, we can consider the following hypothesis tests using F-tests:

- $H_0 : a_i = 0 \quad \forall i$ vs $H_1 : a_i \neq 0$ for at least one $i \rightarrow F_{obs} = MSA/MSAB \sim F_{t_a-1, (t_a-1)(t_b-1)}$
- $H_0 : \sigma_b^2 = 0$ vs $H_1 : \sigma_b^2 > 0 \rightarrow F_{obs} = MSB/MSE \sim F_{t_b-1, t_a t_b (r-1)}$
- $H_0 : \sigma_{ab}^2 = 0$ vs $H_1 : \sigma_{ab}^2 > 0 \rightarrow F_{obs} = MSAB/MSE \sim F_{(a-1)(b-1), t_a t_b (r-1)}$

Now what about the estimates of a_i and b_j ? This is something we typically discuss, but we really have not yet at this point. The key is that we can estimate the fixed effects just as before. However, estimation of the random effects b_j is a bit different and will not be discussed in this class. So the fixed effects are estimated by:

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{...} \\
\hat{a}_i &= \bar{y}_{i..} - \bar{y}_{...} \quad \forall i
\end{aligned}$$

Now we need to estimate the three variance parameters. Again we use the method of moments and set the empirical means equal to the theoretical means under the model:

$$\begin{aligned}
MSE &= \sigma^2 \\
MSB &= \sigma^2 + t_a r \sigma_b^2 \\
MSAB &= \sigma^2 + r \sigma_{ab}^2
\end{aligned}$$

Solving the three equations for the three unknowns we have:

$$\begin{aligned}\hat{\sigma}_2 &= MSE \\ \hat{\sigma}_b^2 &= (MSB - MSE)/(t_a r) \\ \hat{\sigma}_{ab}^2 &= (MSAB - MSE)/(r)\end{aligned}$$

- Back to the example:

```
> ## pos is random
> ## temp is fixed
> rm(list = ls())
>
> y <- c(570, 1063, 565,
+       565, 1080, 510,
+       583, 1043, 590,
+       528, 988, 526,
+       547, 1026, 538,
+       521, 1004, 532)
>
> pos <- as.factor(rep(1:2, each=9))
> temp <- as.factor(rep(c(800,825, 850), 6))
>
> ## anova
> fit.lm <- lm(y ~ as.factor(temp) + as.factor(pos) + as.factor(temp):as.factor(pos))
>
> ## get the Mean Squares
> MSA <- anova(fit.lm)[1,3]
> MSB <- anova(fit.lm)[2,3]
> MSAB <- anova(fit.lm)[3,3]
> MSE <- anova(fit.lm)[4,3]
>
> MSA
[1] 472671.1
> MSB
[1] 7160.056
> MSAB
[1] 409.0556
> MSE
[1] 447.5556
>
> ## calculate the observed F-statistics
> F.A.obs <- MSA/MSAB
> F.B.obs <- MSB/MSE
> F.AB.obs <- MSAB/MSE
>
> F.A.obs
[1] 1155.518
> F.B.obs
[1] 15.99814
> F.AB.obs
[1] 0.9139772
```

```

>
> ## calculate the p-values
> p.value.A <- 1-pf(F.A.obs, 3-1, (3-1)*(2-1))
> p.value.B <- 1-pf(F.B.obs, 2-1, 3*2*(3-1))
> p.value.AB <- 1-pf(F.AB.obs, (3-1)*(2-1), 3*2*(3-1))
>
> p.value.A
[1] 0.0008646645
> p.value.B
[1] 0.001762434
> p.value.AB
[1] 0.4271101

```

- Now lets get the estimates for the various fixed effects:

```

> mu.hat
[1] 709.9444
> a.i.hat
      800      825      850
-157.6111  324.0556 -166.4444

```

- Now lets get the estimates for the various variances:

```

> ## variance estimates
> sigma.sq.hat <- MSE
> sigma.sq.B.hat <- (MSB-MSE)/(3*3)
> sigma.sq.AB.hat <- (MSAB-MSE)/(3)
>
> sigma.sq.hat
[1] 447.5556
> sigma.sq.B.hat
[1] 745.8333
> sigma.sq.AB.hat
[1] -12.83333

```

Notice that the variance in $y_{ijk} \rightarrow V(y_{i,j,k}) = \sigma_b^2 + \sigma_{ab}^2 + \sigma^2$ is dominated by the between variance of the positions σ_b^2 . Also, the variance on the interaction is negative. Computationally this can happen, and it we should set $\sigma_{ab}^2 = 0$, which is exactly what is suggested by the p-value for the hypothesis test.