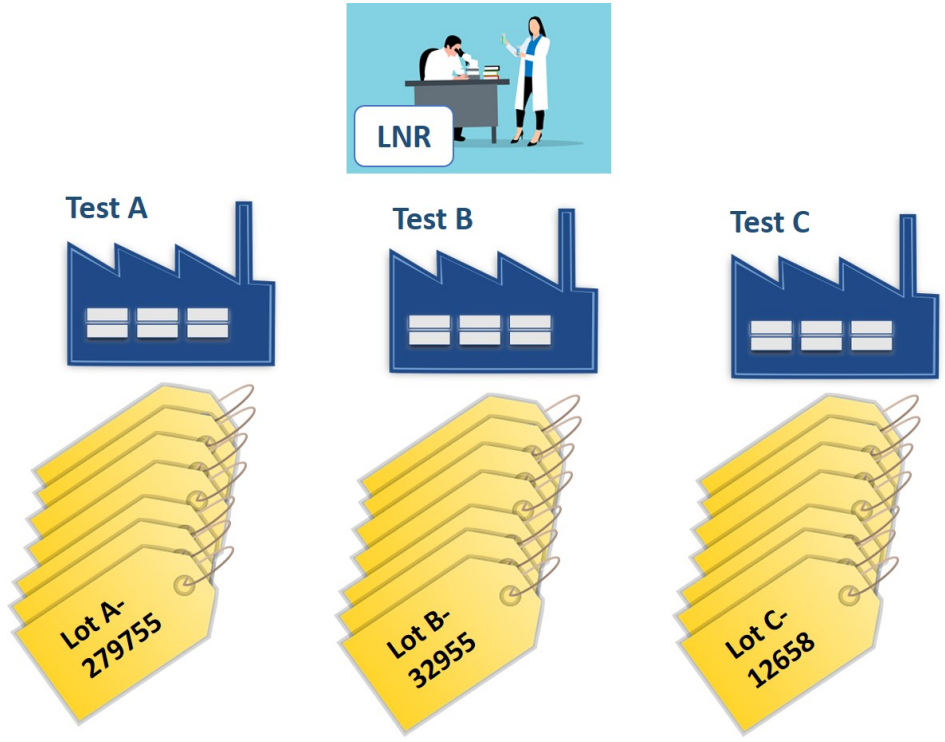
Introduction aux modèles linéaires mixtes

Marie Laure Delignette-Muller - VetAgro Sup

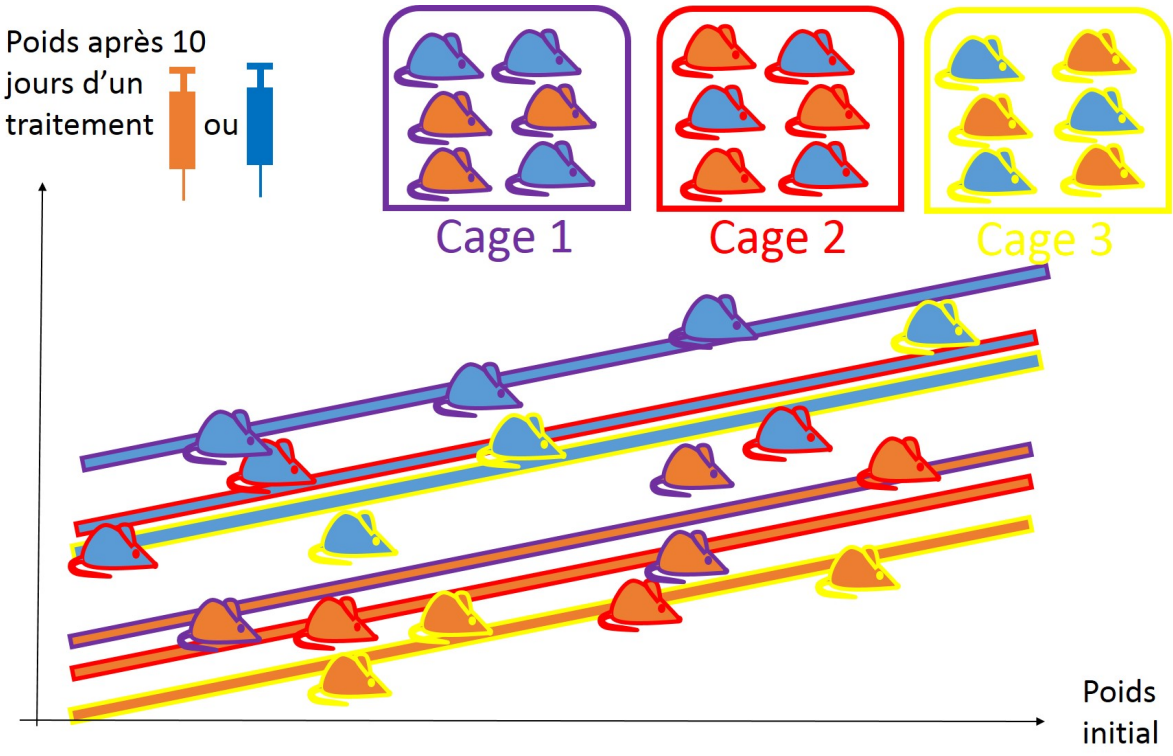


Introduction et définitions

C'est quoi un **modèle mixte** ?

Un **modèle linéaire est dit mixte** dès qu'il prend en compte **au moins un facteur aléatoire**.

Ex. : modélisation d'une variable quantitative poids 10 jrs ap. traitement) en fonction de deux variables explicatives (poids initial et traitement) en prenant en compte un facteur aléatoire (cage).



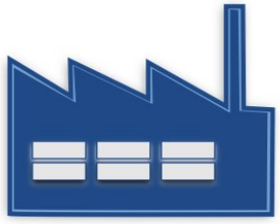
Notre exemple fil rouge

Nous présenterons les concepts de base et divers modèles simples sur une problématique d'évaluation de tests Elisa utilisés pour le diagnosic de la fièvre Q chez les caprins.

La variable mesurée, un rapport de densité optique (par rapport à un témoin), est quantitative continue. En pratique c'est en fonction de la position de la mesure par rapport à un seuil fixé par le fabricant du test que le diagnostic est établi (positif ou négatif).

Mais en ce qui nous concerne, nous étudierons l'**impact des différentes sources de variabilité** sur le **rapport de densité optique qu'on appellera ODR** et qui est une **variable quantitative continue**.

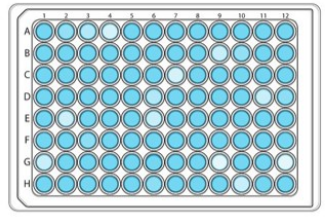
Sources de variabilité dans notre exemple



Fabricant du test 3 tests disponibles : A, B, C

Lot de fabrication de chaque kit utilisé

Laboratoire qui réalise le test



De la plaque sur laquelle le test est réalisé

L'ODR, pour un échantillon biologique donné, dépend :

- du **test** Elisa (3 tests commercialisés),
- du **lot** de fabrication,
- du **laboratoire** qui réalise le test et
- de la **plaque** sur laquelle le test est réalisé.

Définition d'un facteur fixe

Un **facteur** est considéré comme **fixe** si toutes les modalités de ce facteur sont testées dans l'expérience (ex. : facteur "traitement", "sexe", ...)

L'effet de ce facteur est supposé **prévisible** d'une expérimentation à l'autre (effet = différence entre les moyennes pour les modalités de ce facteur).

Une variable explicative quantitative (ex. : "âge", "temps", ...) est considérée comme fixe (effet = coefficient de régression).

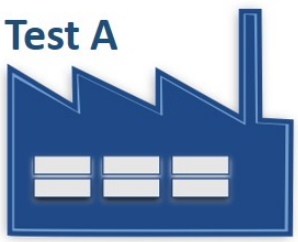
Modèle fixe d'ANOVA1 sur notre exemple (1/2)

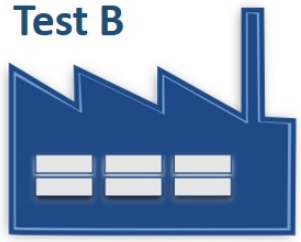
Comparaison des 3 tests commercialisés par le laboratoire de référence

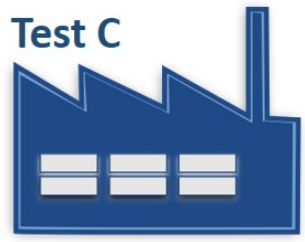
On veut comparer les 3 tests disponibles. On fait, au labo de référence, 20 mesures indépendantes avec chacun des tests, à partir de 60 échantillons d'un même serum répartis aléatoirement entre les 3 tests.

On suppose que la variabilité de la mesure est la même pour les 3 tests (à vérifier *a posteriori*).

Modèle fixe d'ANOVA1 sur notre exemple (2/2)







Facteur A fixe (facteur **test** avec 3 modalités)

$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$

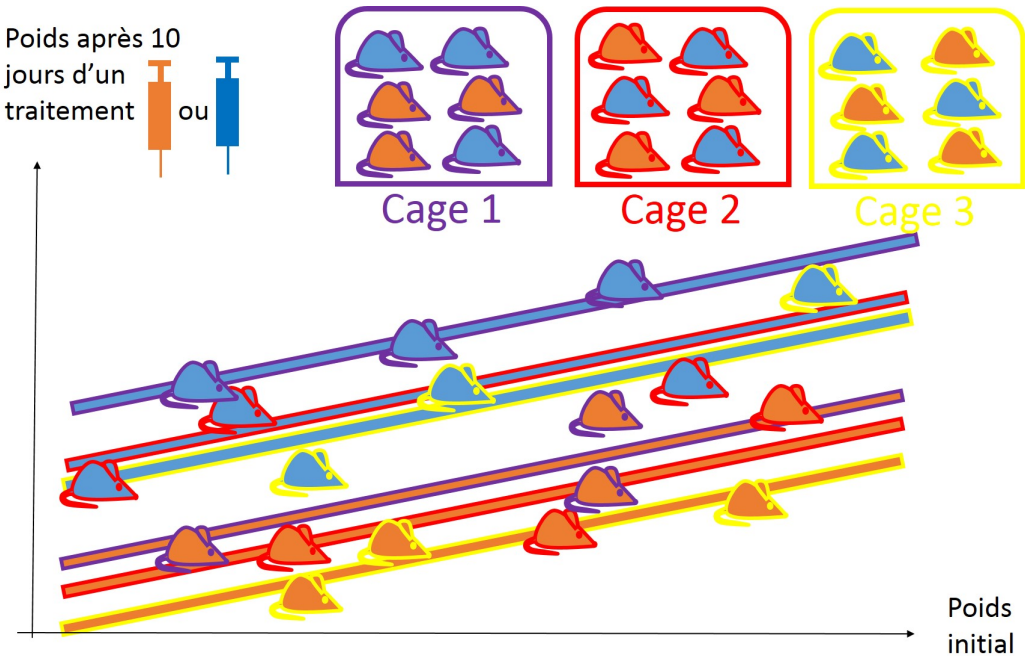
avec $\epsilon_{ij} \sim N(0, \sigma)$

Hypothèse nulle testée : $\sum \alpha_i^2 = 0$

Définition d'un facteur aléatoire

Un **facteur** est considéré comme **aléatoire** si seul un échantillon aléatoire des modalités du facteur est testé dans l'expérience (ex. : facteur "animal", "cage", ...)

L'effet de ce facteur est supposé **imprévisible** d'une expérimentation à l'autre.



Dans l'exemple ci-dessus, le traitement et le poids sont des facteurs fixes et la cage est un facteur aléatoire.

Modèle aléatoire d'ANOVA1 sur notre exemple (1/2)

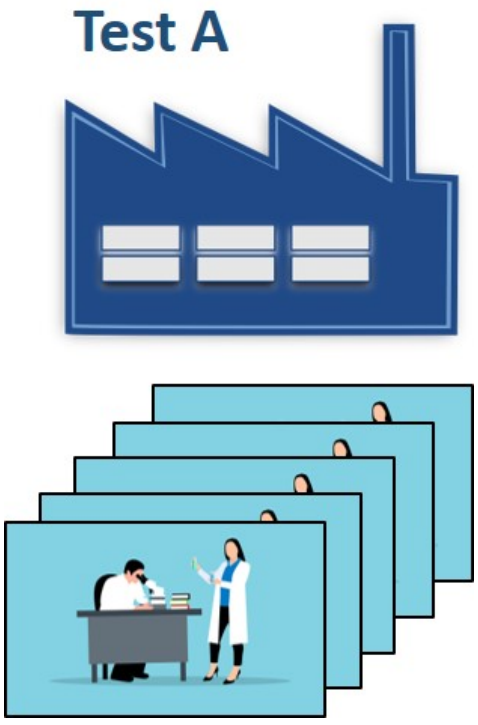
Essai interlaboratoire

On veut évaluer la variabilité interlaboratoire pour un test Elisa donné. On prend au hasard 15 laboratoires, et on envoie à chaque laboratoire 10 échantillons d'un même serum (donc 150 échantillons au total sont répartis aléatoirement entre les 15 laboratoires).

On suppose que la variabilité intralaboratoire est la même dans tous les labos (à vérifier *a posteriori*).

On suppose aussi, souvent sans le dire, que les moyennes des laboratoires sont réparties suivant une loi normale (à vérifier *a posteriori* aussi !).

Modèle aléatoire d'ANOVA1 sur notre exemple (2/2)



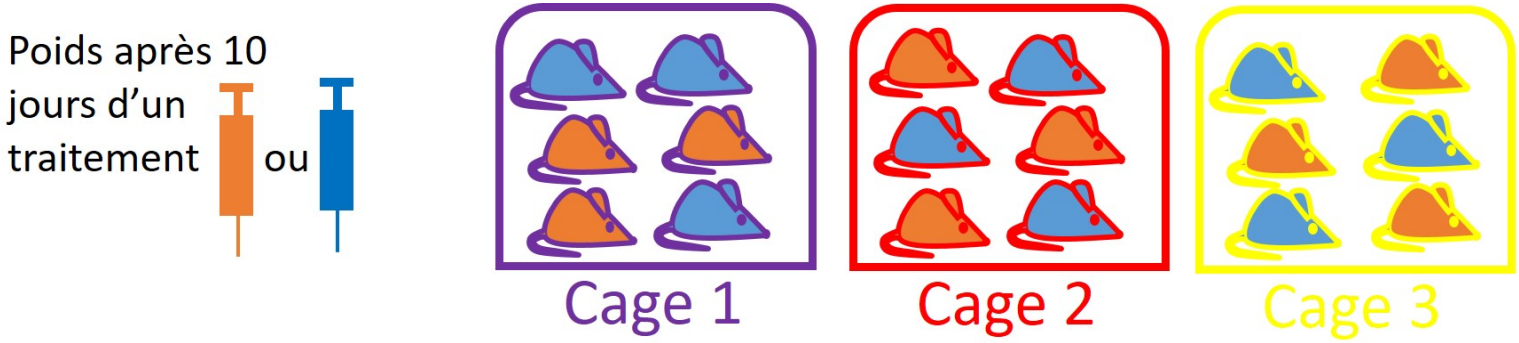
Facteur A aléatoire (facteur **laboratoire**)

$X_{ij} = \mu + A_i + \epsilon_{ij}$ avec $\epsilon_{ij} \sim N(0, \sigma)$ et $A_i \sim N(0, \sigma_A)$

Ce qui nous intéresse est l'estimation de la variabilité entre les laboratoires (caractérisée par l'écart type σ_A)

Définition de facteurs croisés

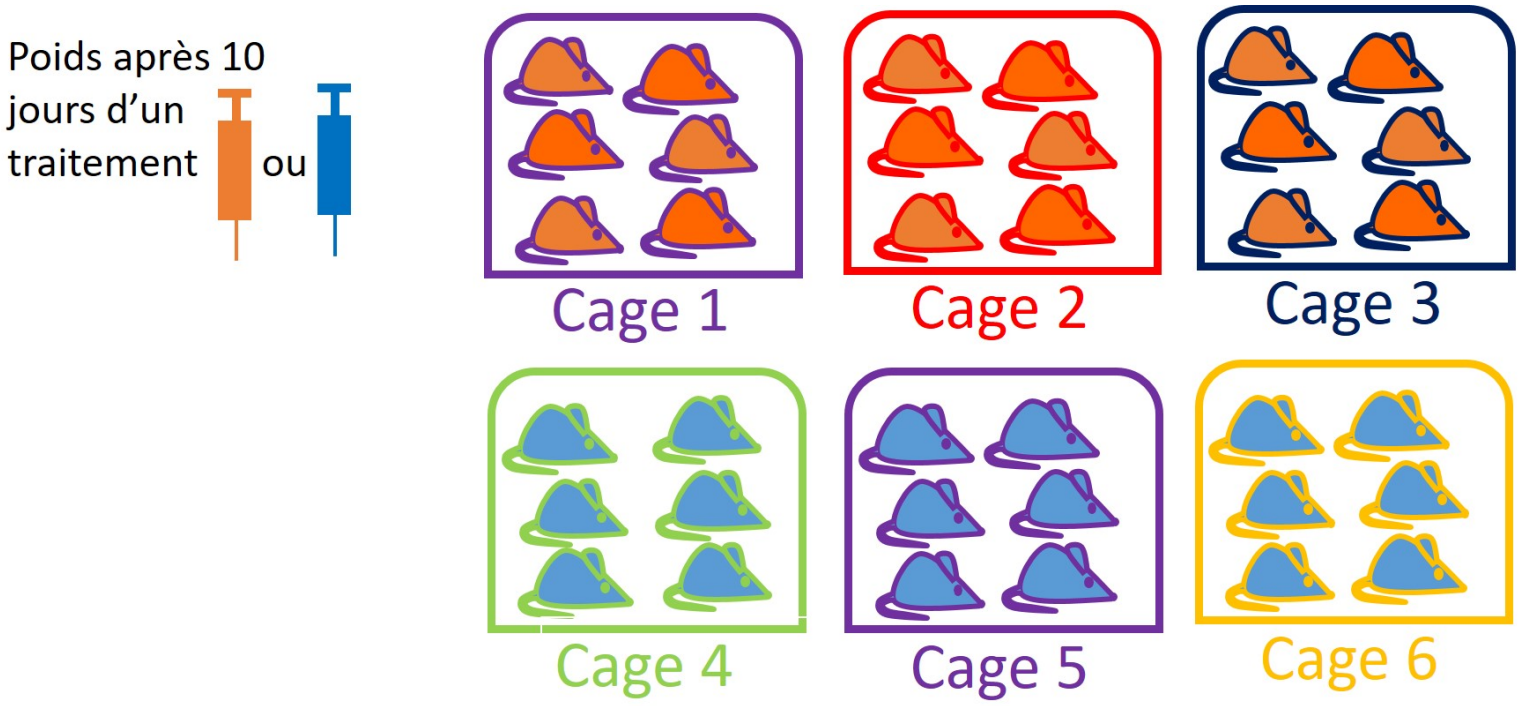
Deux facteurs A et B sont dits croisés si toutes les modalités testées de B sont croisées avec toutes les modalités testées de A.



Dans l'exemple ci-dessus, le **traitement** et la **cage** sont croisés.

Définition de facteurs hiérarchisés (ou imbriqués)

Deux facteurs A et B sont dits hiérarchisés, avec B imbriqué dans A, si les modalités possibles de B sont dépendantes de la modalité de A (chaque modalité de B ne peut être associée qu'à une seule modalité de A).



Dans l'exemple ci-dessus, le **traitement** et la **cage** sont hiérarchisés : la cage est imbriquée dans le traitement.

Objectifs pédagogiques

- Savoir caractériser le modèle correspondant à un plan d'expérience donné (identification et nombre de facteurs fixes et aléatoires, hiérarchie éventuelle entre facteurs, ...),
- comprendre le principe des modèles mixtes à partir de l'étude de cas simples à 2 facteurs,
- savoir réaliser (avec le package R `lme4`) et interpréter l'analyse des données dans le cas de 2 facteurs,
- savoir discuter avec un statisticien dans le cadre de la mise en place d'un plan d'expérience à plus de 2 facteurs et de l'analyse de ses données.

Modèle aléatoire à un facteur

Visualisation des données sur l'exemple proposé.

```
str(d_a)
```

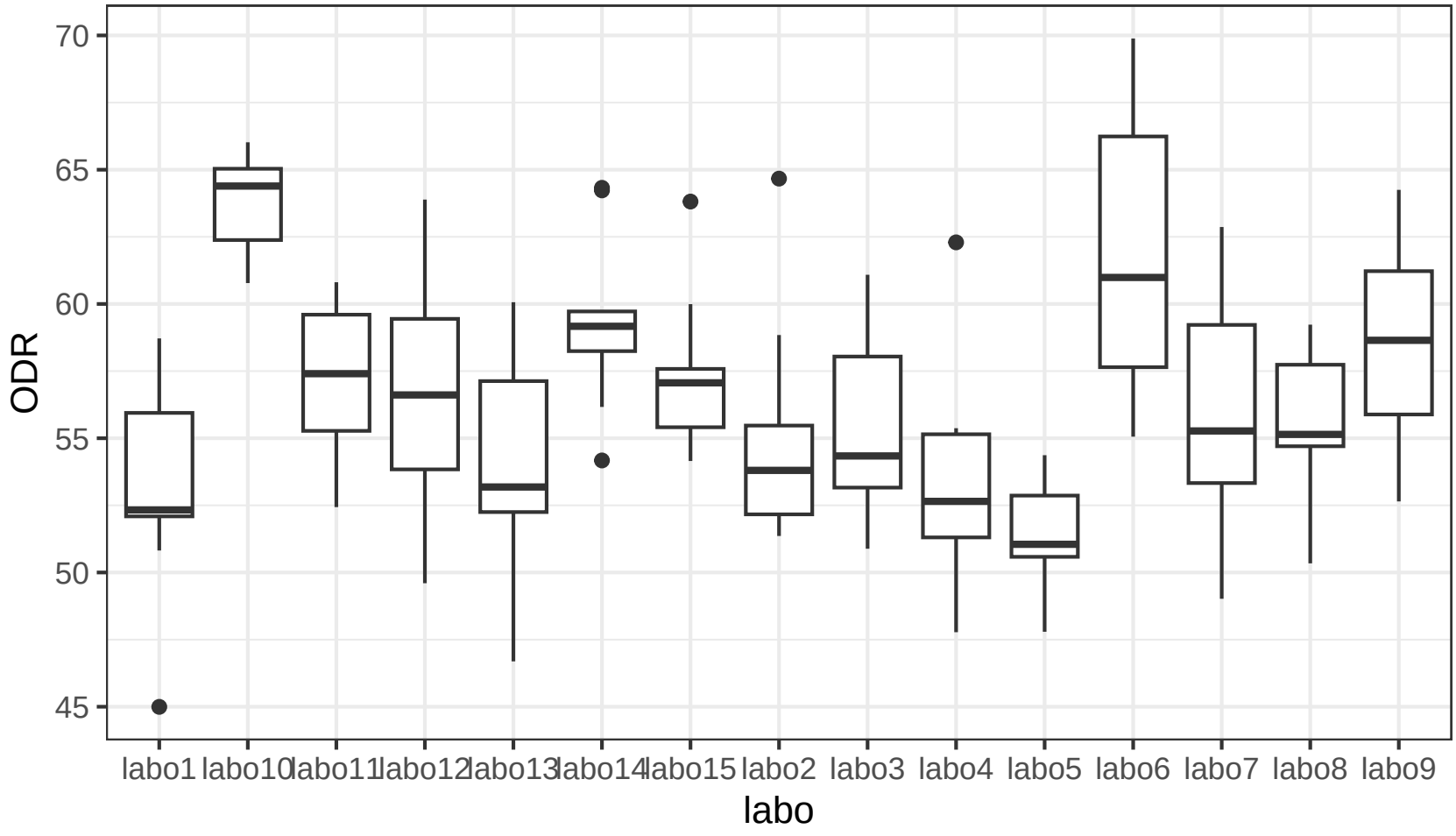
```
## 'data.frame':    150 obs. of  2 variables:
##  $ labo: Factor w/ 15 levels "labo1","labo10",..: 1 1 1 1 1 1
##  $ ODR : num  55.5 58.7 45 56.1 56.4 ...
```

```
xtabs(~ labo, data = d_a)
```

```
## labo
##  labo1 labo10 labo11 labo12 labo13 labo14 labo15  labo2  labo
##     10     10     10     10     10     10     10     10      1
##  labo6  labo7  labo8  labo9
##     10     10     10     10
```

Une représentation graphique des données

```
ggplot(data = d_a, aes(x = labo, y = ODR)) + geom_boxplot()
```



Ajustement du modèle aux données

```
(mm_a <- lmer(ODR ~ (1|labo), data = d_a))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ (1 | labo)
##    Data: d_a
## REML criterion at convergence: 840
## Random effects:
##  Groups   Name        Std.Dev.
##  labo     (Intercept) 3.10
##  Residual             3.61
## Number of obs: 150, groups:  labo, 15
## Fixed Effects:
## (Intercept)
##        56.6
```

Estimations : $\hat{\mu} = 56.6$, $\hat{\sigma} = 3.61$ **et** $\hat{\sigma}_{labo} = 3.10$

Intervalles de confiance

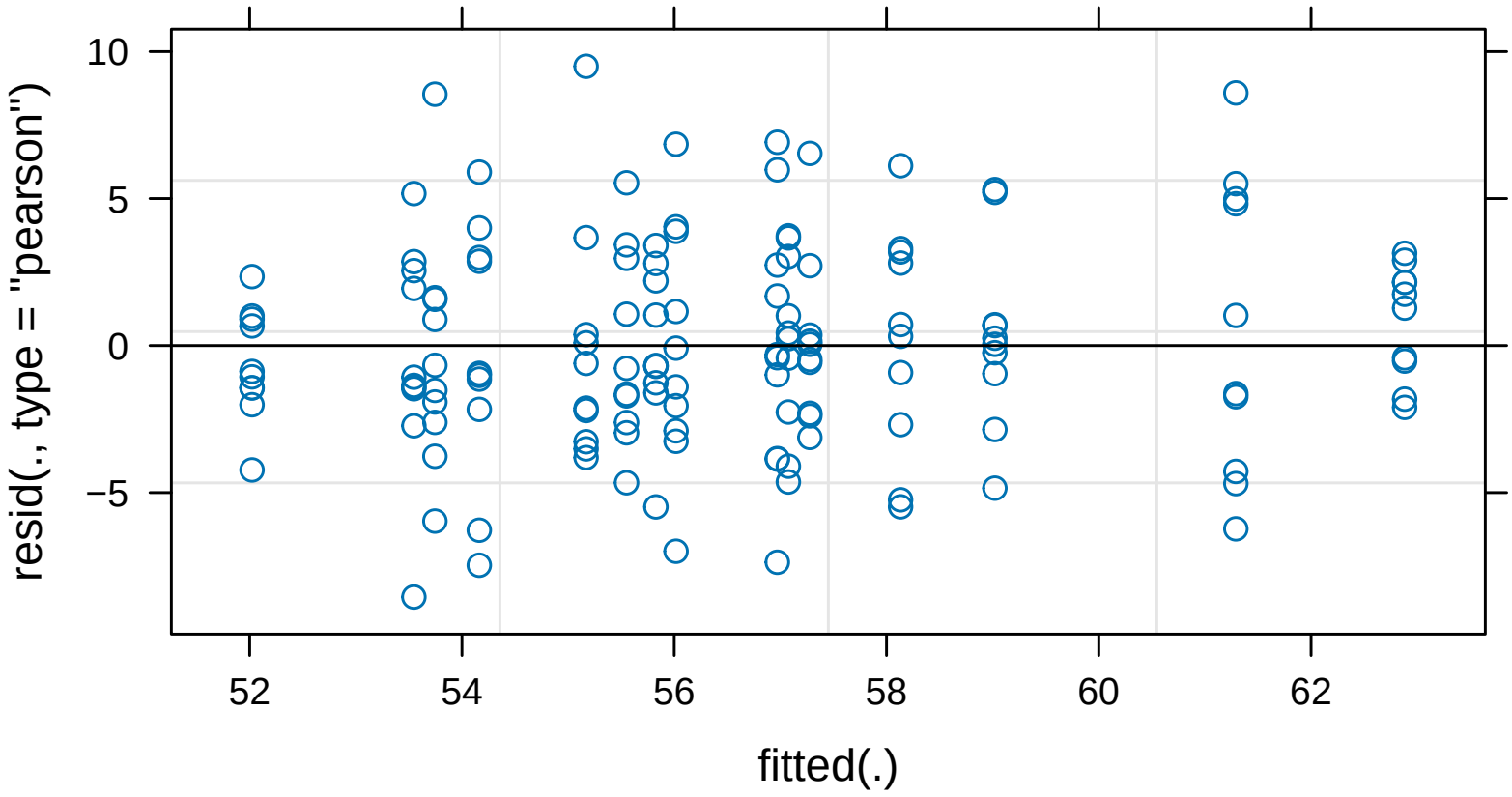
```
confint(mm_a)
```

```
##                2.5 % 97.5 %
## .sig01          2.02   4.66
## .sigma          3.21   4.08
## (Intercept) 54.85  58.31
```

Celui appelé `sigma` correspond toujours à l'écart type résiduel $\hat{\sigma}$.

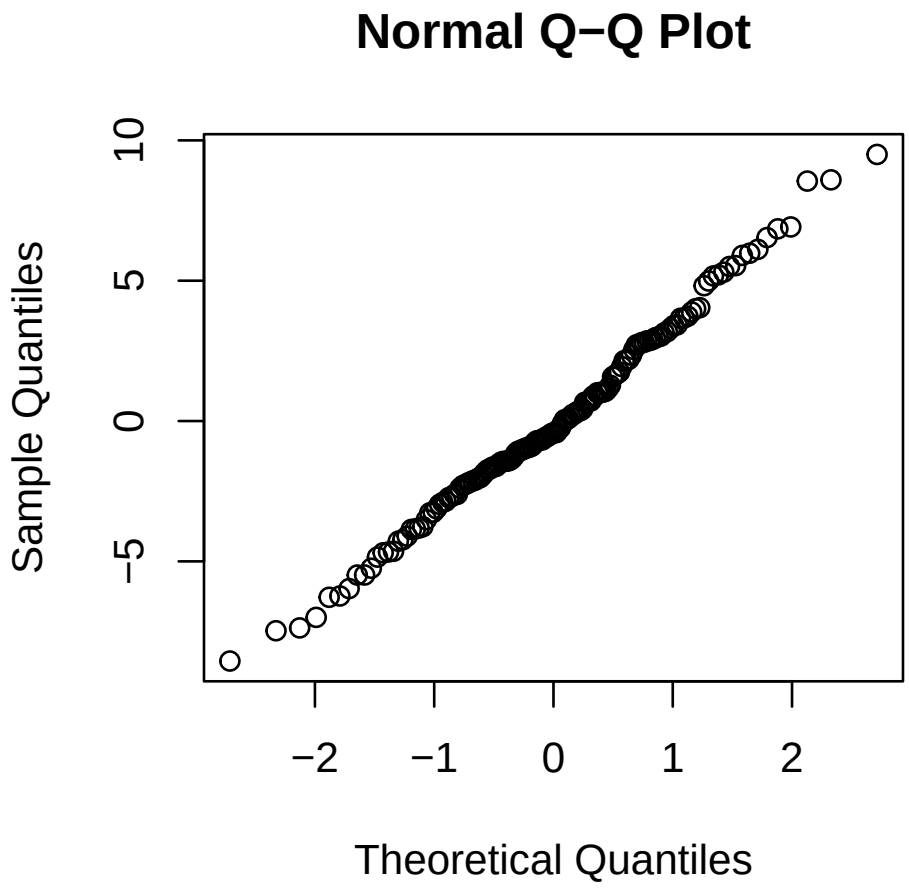
Vérification des conditions d'utilisation - graphe des résidus

```
plot(mm_a)
```



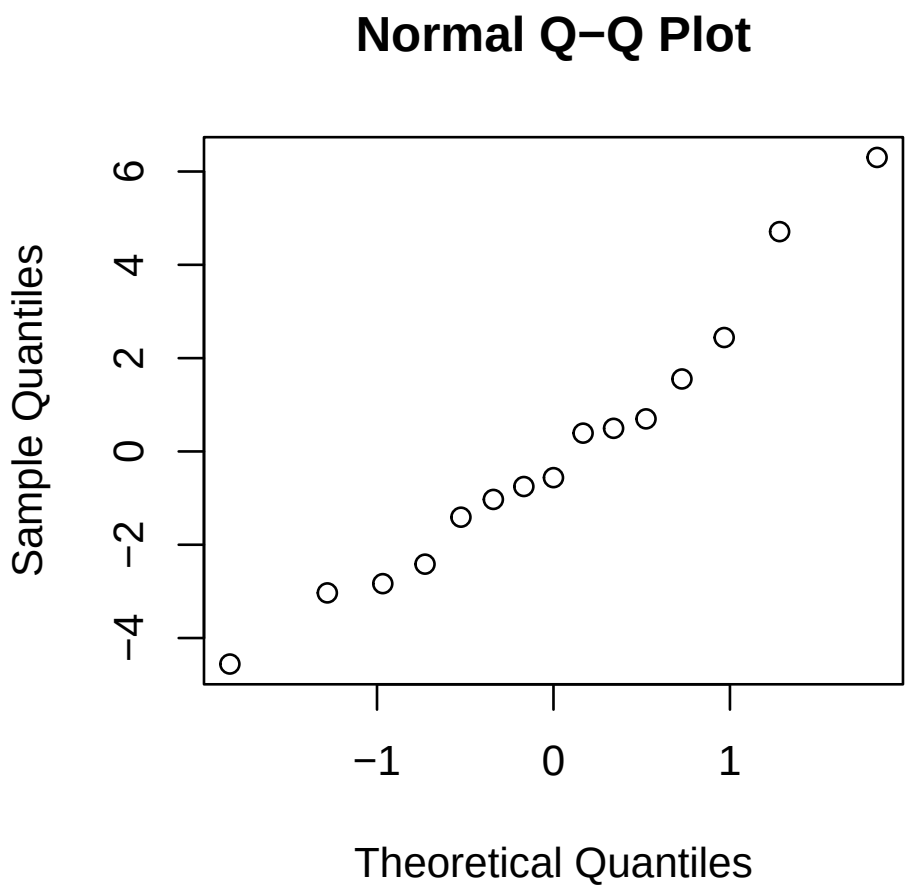
Vérification des conditions d'utilisation - diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm_a))
```



Vérification des conditions d'utilisation - diagramme quantile-quantile des effets aléatoires

```
qqnorm(ranef(mm_a)$labo[, 1])
```



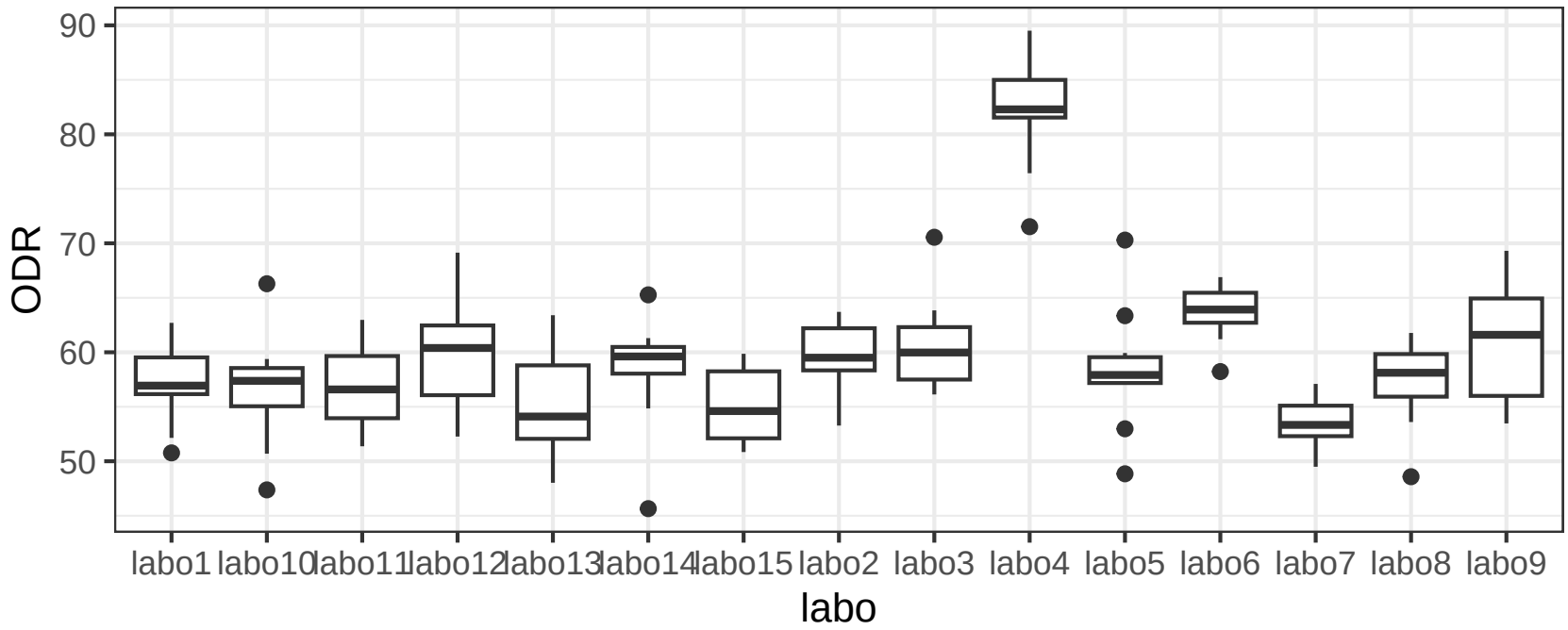
A votre tour

Utilisez la même démarche pour analyser un autre jeu de données correspondant à la même expérience : fichier “data_Elisa_labo.txt”.

Quelle condition d’utilisation du modèle vous semble non respectée ?

Import et examen des données

```
d_a_bis <- read.table("DATA/data_Elisa_labo.txt",
                      header = TRUE, stringsAsFactors = TRUE)
ggplot(data = d_a_bis, aes(x = labo, y = ODR)) + geom_boxplot()
```



Un des labos semble se démarquer des autres !

Ajustement du modèle

```
(mm_a_bis <- lmer(ODR ~ (1|labo), data = d_a_bis))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ (1 | labo)
##    Data: d_a_bis
## REML criterion at convergence: 916
## Random effects:
##  Groups   Name        Std.Dev.
##  labo     (Intercept) 6.63
##  Residual             4.43
## Number of obs: 150, groups:  labo, 15
## Fixed Effects:
## (Intercept)
##        59.7
```

Graphe des résidus

```
plot(mm_a_bis)
```

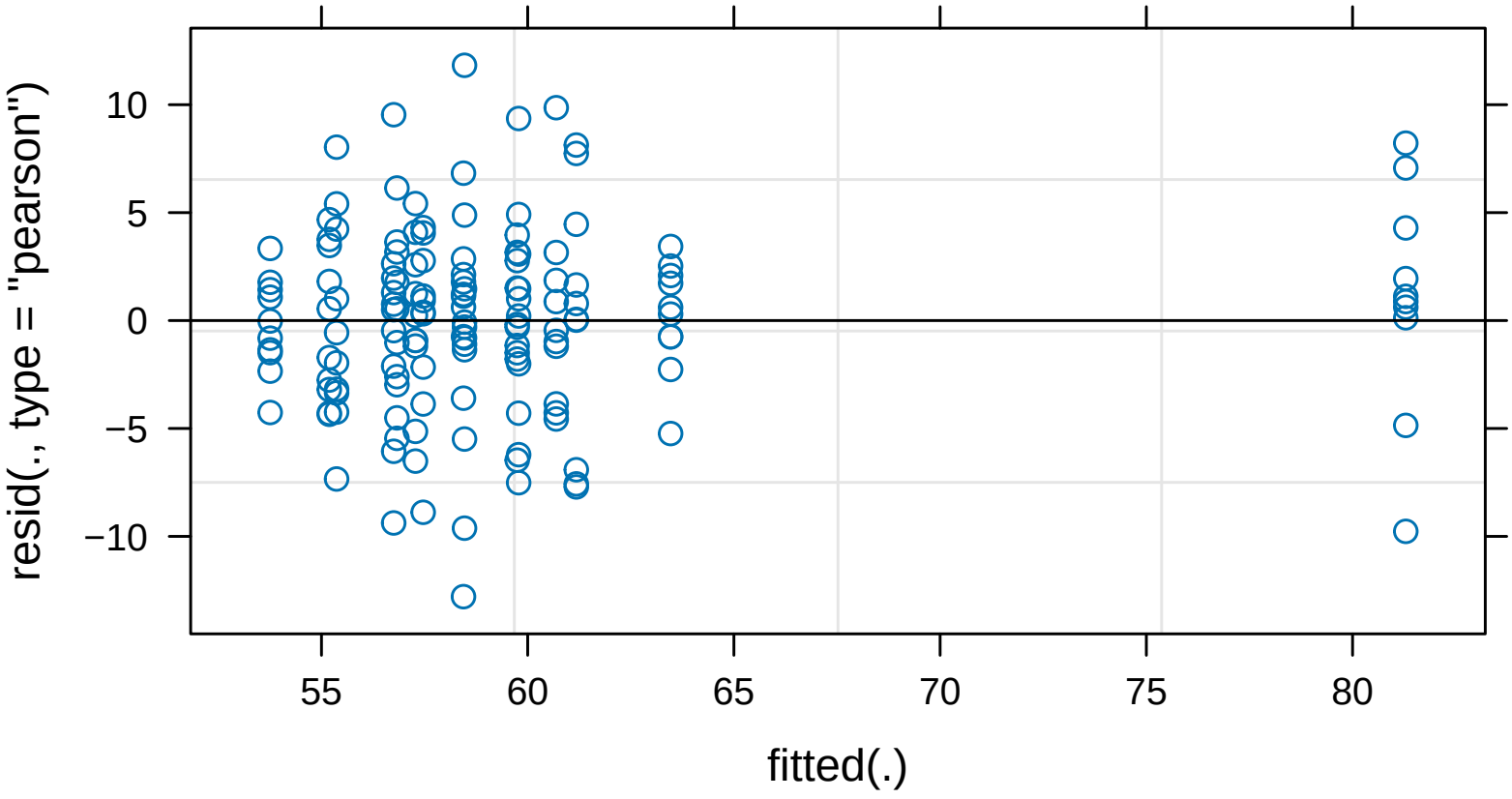


Diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm_a_bis))
```

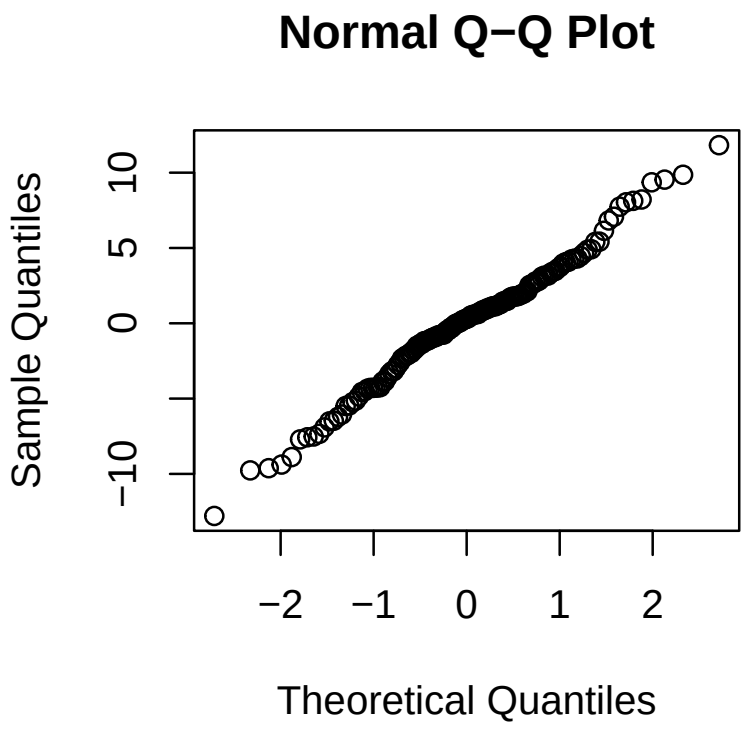
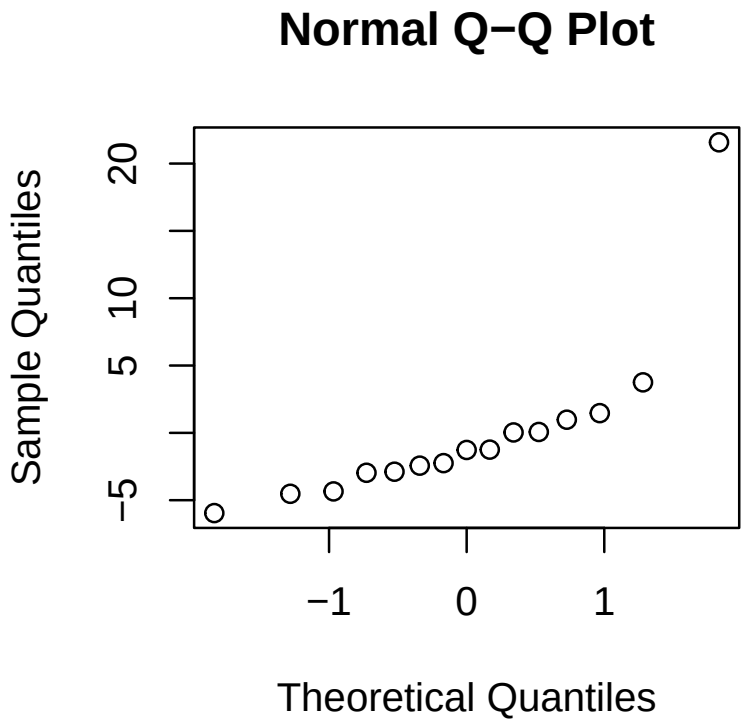


Diagramme quantile-quantile des effets aléatoires

```
qqnorm(ranef(mm_a_bis)$labo[, 1])
```

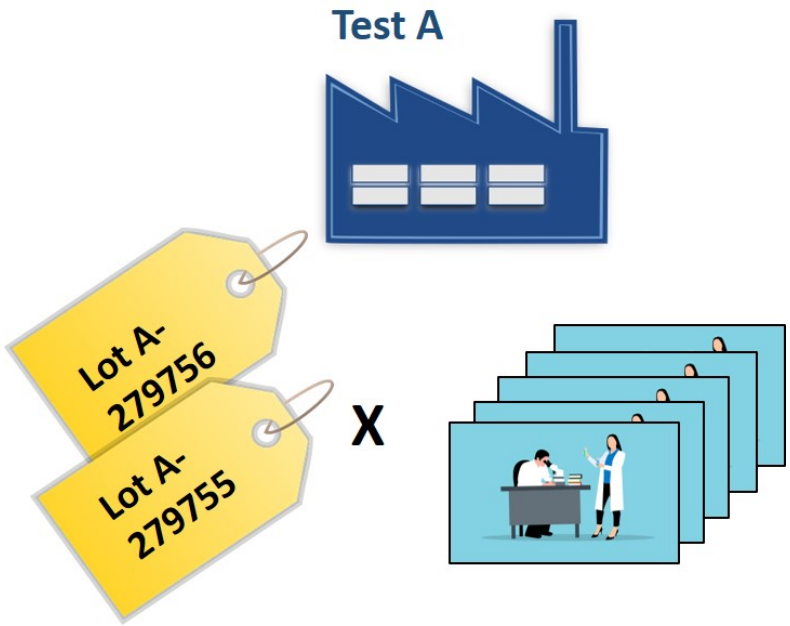


L'hypothèse de distribution normale des effets aléatoires est clairement remise en question à partir de ce graphe, du fait du labo qui répond très différemment des autres.

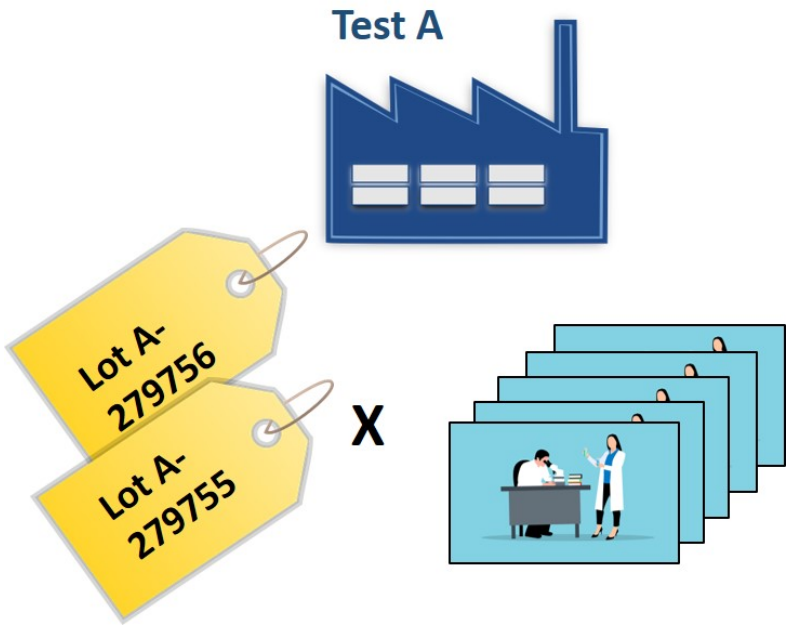
Modèle croisé mixte à 2 facteurs

Exemple

Dans un essai interlaboratoire, on veut comparer deux lots de fabrication. On demande à chaque laboratoire de faire 10 mesures avec le lot 1 du test, 10 mesures avec le lot 2. Les 2 lots sont les mêmes pour tous les laboratoires. On envoie à chacun des 15 laboratoires tirés au hasard, 20 échantillons d'un même serum (donc 300 échantillons au total sont répartis aléatoirement entre les 15 laboratoires, et pour chaque laboratoire les 20 échantillons sont répartis aléatoirement sur les 2 lots du test).



Facteurs



Deux facteurs croisés à prendre en compte :

- **laboratoire (aléatoire)**
- **lot (fixe)** ici car on s'intéresse juste à la différence entre ces deux lots. Si on avait voulu caractériser plus généralement la variabilité entre les lots, on aurait pris plus de deux lots dans l'expérience.

Modèle théorique

Deux modèles mixtes peuvent en théorie être utilisés dans un tel cas, un modèle complet et un modèle simplifié négligeant l'interaction entre le facteur fixe (A) et le facteur aléatoire (B) (c'est-à-dire la potentielle variabilité de l'effet fixe - différence entre les 2 lots - due au facteur aléatoire - le laboratoire).

- Modèle complet :
  $X_{ijk} = \mu + \alpha_i + B_j + AB_{ij} + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$,
  $B_j \sim N(0, \sigma_B)$ et $AB_{ij} \sim N(0, \sigma_{AB})$
- Modèle simplifié sans interaction (plus courant) :
  $X_{ijk} = \mu + \alpha_i + B_j + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$ et $B_j \sim N(0, \sigma_B)$

Visualisation des données

```
str(d_b)
```

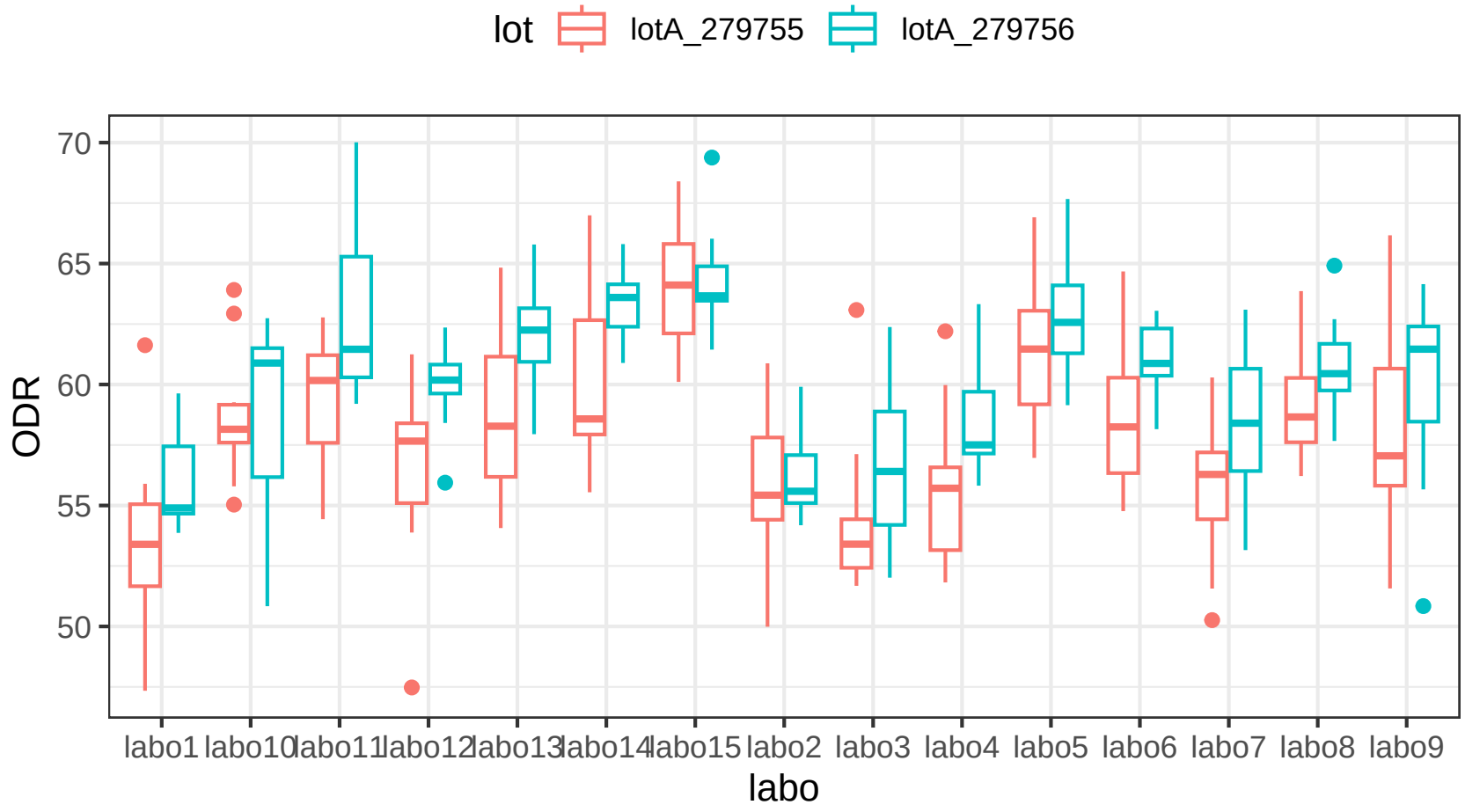
```
## 'data.frame':    300 obs. of  3 variables:
##  $ labo: Factor w/ 15 levels "labo1","labo10",..: 1 1 1 1 1 1
##  $ lot : Factor w/ 2 levels "lotA_279755",..: 1 2 1 2 1 2 1 2
##  $ ODR : num  55.2 59.6 47.3 57.7 55.9 ...
```

```
xtabs(~ labo + lot, data = d_b)
```

```
##          lot
## labo      lotA_279755 lotA_279756
##   labo1            10          10
##   labo10           10          10
##   labo11           10          10
##   labo12           10          10
##   labo13           10          10
##   labo14           10          10
##   labo15           10          10
##   labo2            10          10
##   labo3            10          10
##   labo4            10          10
```

Une représentation graphique des données

```
ggplot(data = d_b, aes(x = labo, col = lot, y = ODR)) +
  geom_boxplot() + theme(legend.position = "top")
```



Ajustement du modèle sans interaction aux données

```
(mm_b <- lmer(ODR ~ lot + (1|labo), data = d_b))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ lot + (1 | labo)
##    Data: d_b
## REML criterion at convergence: 1554
## Random effects:
##  Groups   Name        Std.Dev.
##  labo     (Intercept) 2.58
##  Residual             3.02
## Number of obs: 300, groups:  labo, 15
## Fixed Effects:
##    (Intercept)  lotlotA_279756
##          58.03            2.09
```

- **Estimation des effets fixes** : $\hat{\mu}_{lotA_{279755}} = 58.03$ et $\hat{\mu}_{lotA_{279756}} = 58.03 + 2.09$
- **Estimation des effets aléatoires** : $\hat{\sigma} = 3.02$ et $\hat{\sigma}_{labo} = 2.58$

Intervalles de confiance

```
confint(mm_b)
```

```
##                   2.5 % 97.5 %
## .sig01             1.75   3.81
## .sigma             2.79   3.28
## (Intercept)       56.60  59.45
## lotlotA_279756     1.41   2.78
```

Celui appelé `sigma` correspond toujours à l'écart type résiduel $\hat{\sigma}$.

Graphe des résidus

```
plot(mm_b)
```

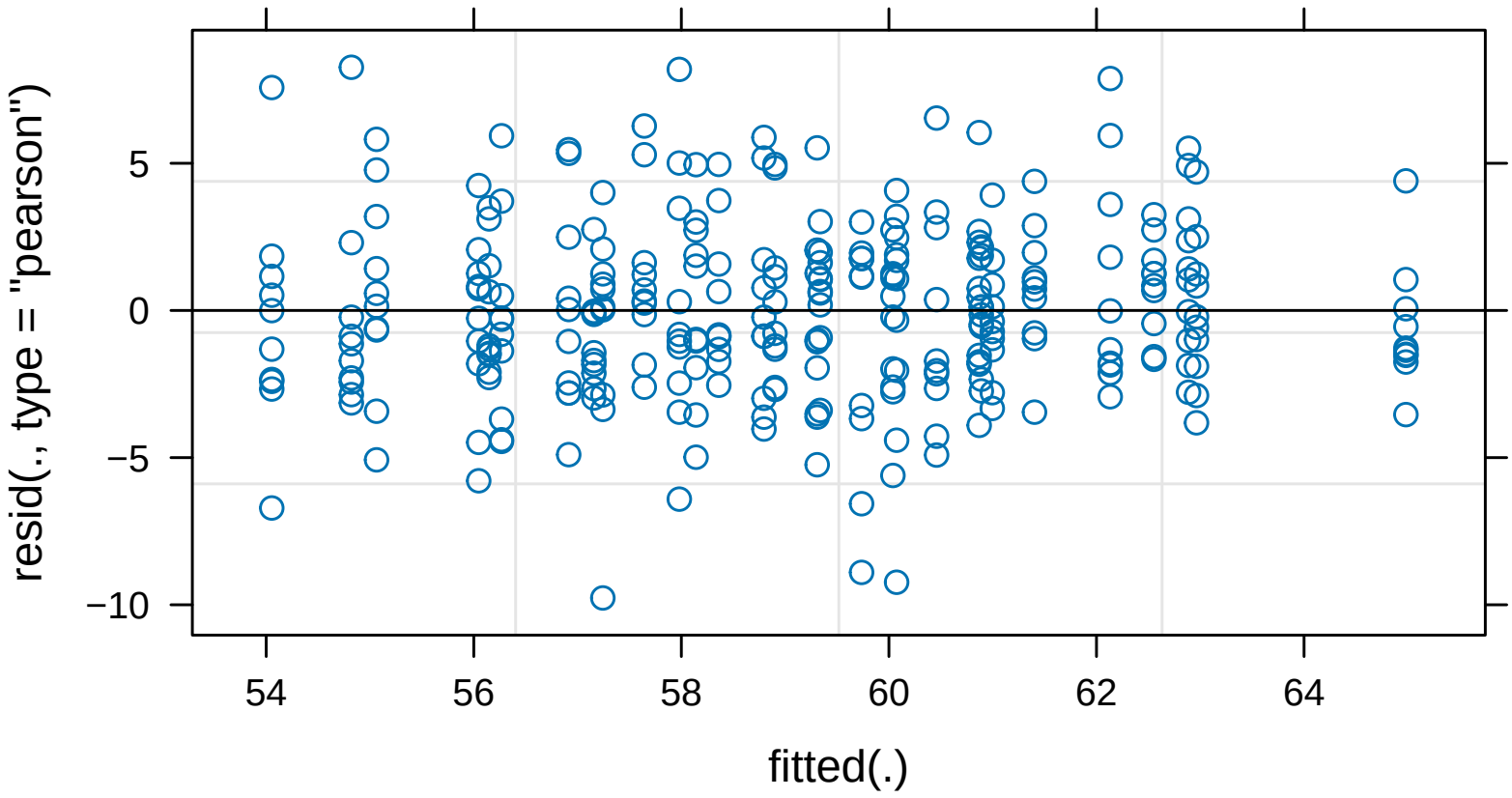


Diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm_b))
```

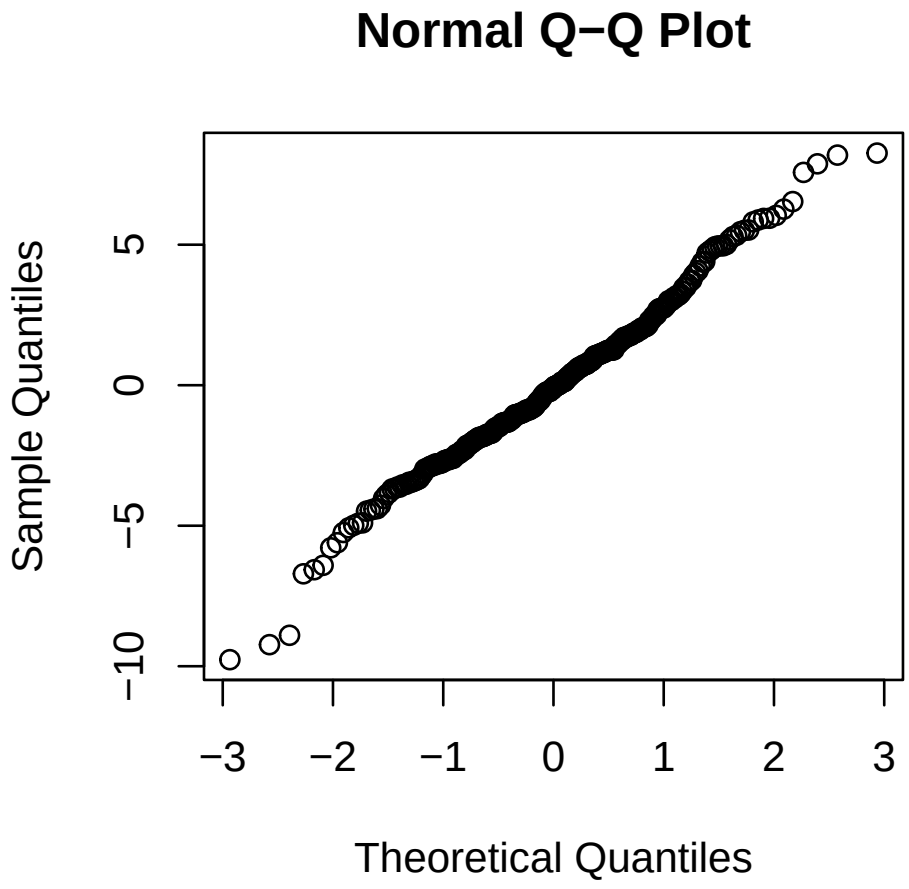
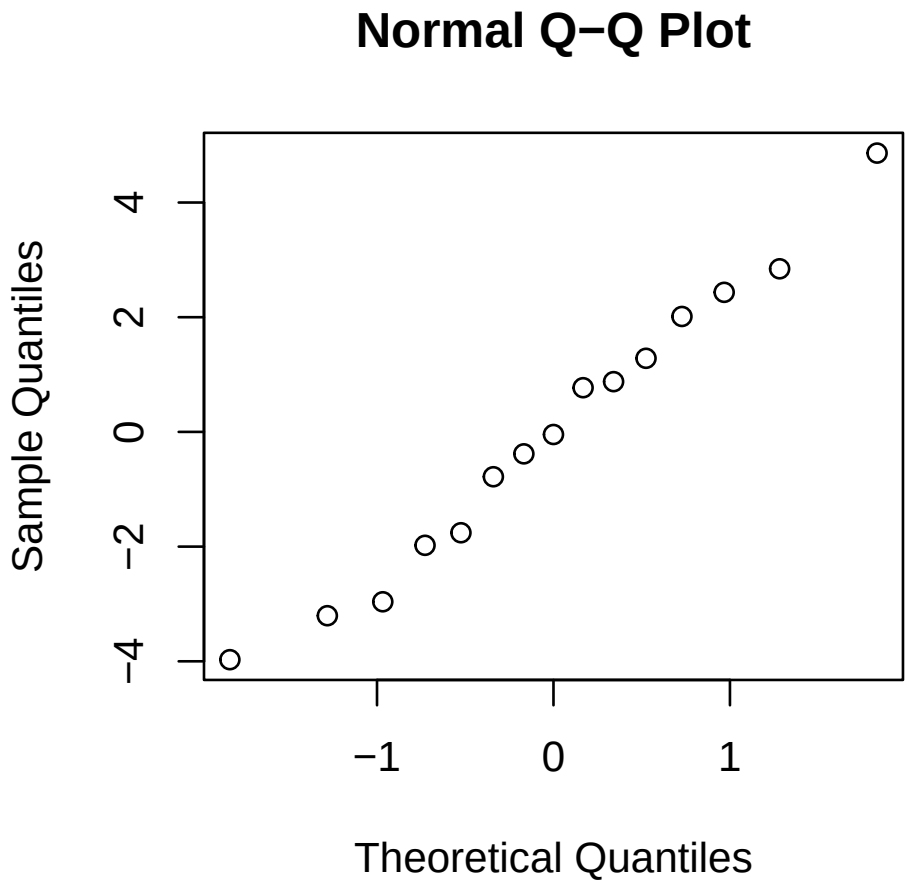


Diagramme quantile-quantile des effets aléatoires

```
qqnorm(ranef(mm_b)$labo[, 1])
```



Modèle avec ou sans interaction ?

Le modèle avec interaction s'écrirait :

```
(mm_b_avec_interaction <-
   lmer(ODR ~ lot + (lot|labo), data = d_b))
```

Il décrirait en plus un effet aléatoire du **labo** sur l'effet **lot**, c'est-à-dire une variabilité entre labos de la différence des moyennes entre les deux labos.

Une telle interaction n'aurait ici pas de fondement biologique évident, mais elle peut être utile dans certains cas.

Lorsqu'il n'y a pas de répétitions (ex. ici si chaque labo ne reçoit que 2 échantillons, un pour chaque lot), il n'est pas possible d'estimer un terme d'interaction et un terme résiduel : seul le modèle sans interaction est utilisable.

Sortie du modèle avec interaction par curiosité

```
(mm_b_avec_interaction <-
   lmer(ODR ~ lot + (lot|labo), data = d_b))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ lot + (lot | labo)
##    Data: d_b
## REML criterion at convergence: 1554
## Random effects:
##  Groups   Name            Std.Dev. Corr
##  labo     (Intercept)     2.639
##           lotlotA_279756  0.115    -1.00
##  Residual                 3.023
## Number of obs: 300, groups:  labo, 15
## Fixed Effects:
##    (Intercept)  lotlotA_279756
##          58.03            2.09
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer
```

Ici l'écart type de l'effet **labo** sur l'effet **lot** apparaît faible (0.115)

A votre tour

Analysez un jeu de données correspondant à la même expérience mais sans répétition : fichier “data_Elisa_labo_lot.txt”.

Import et examen des données

```
d_b_bis <- read.table("DATA/data_Elisa_labo_lot.txt",
                      header = TRUE, stringsAsFactors = TRUE)
str(d_b_bis)
```

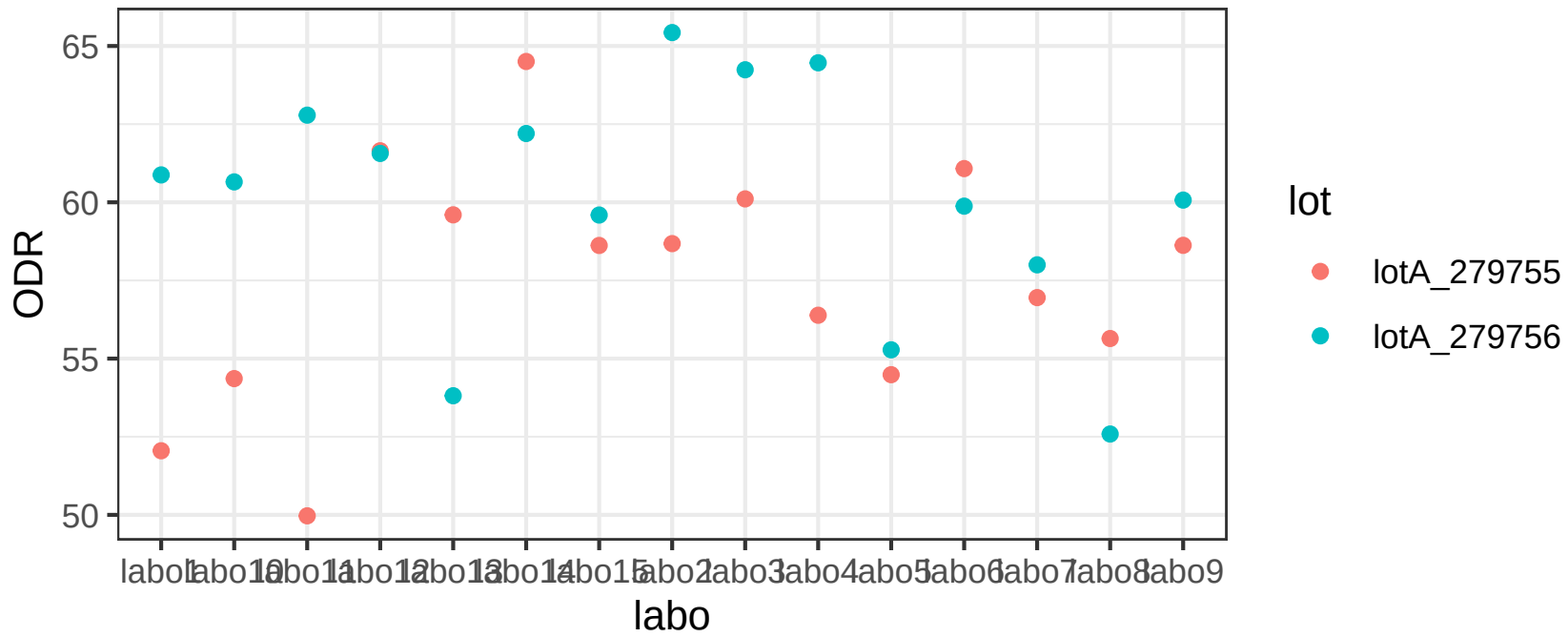
```
## 'data.frame':    30 obs. of  3 variables:
##  $ labo: Factor w/ 15 levels "labo1","labo10",..: 1 1 8 8 9 9
##  $ lot : Factor w/ 2 levels "lotA_279755",..: 1 2 1 2 1 2 1 2
##  $ ODR : num  52.1 60.9 58.7 65.4 60.1 ...
```

```
xtabs(~ labo + lot, data = d_b_bis)
```

```
##         lot
## labo     lotA_279755 lotA_279756
##   labo1              1           1
##   labo10             1           1
##   labo11             1           1
##   labo12             1           1
##   labo13             1           1
##   labo14             1           1
##   labo15             1           1
##   labo2              1           1
```

Représentation des données

```
ggplot(data = d_b_bis, aes(x = labo, y = ODR, col = lot)) +
  geom_point()
```



Ajustement du modèle sans interaction

```
(mm_b_bis <- lmer(ODR ~ lot + (1|labo), data = d_b_bis))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ lot + (1 | labo)
##    Data: d_b_bis
## REML criterion at convergence: 160
## Random effects:
##  Groups   Name        Std.Dev.
##  labo     (Intercept) 1.25
##  Residual             3.61
## Number of obs: 30, groups:  labo, 15
## Fixed Effects:
##    (Intercept)  lotlotA_279756
##          57.51            2.58
```

Message d'erreur en cas d'essai d'ajustement du modèle avec interaction

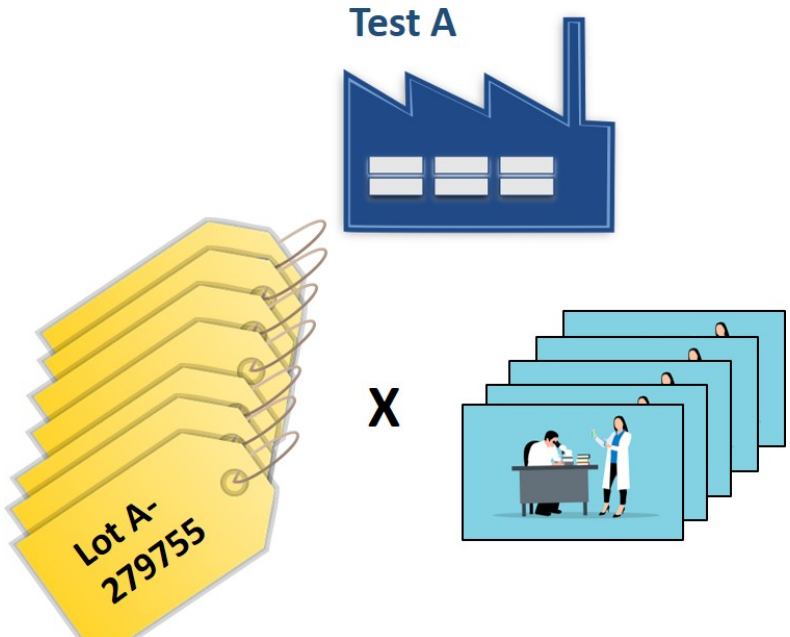
```
(mm_b_bis_avec_interaction <-
    lmer(ODR ~ lot + (lot|labo), data = d_b_bis))
```

```
Error : number of observations (=30) <= number of random
effects (=30) for term (lot | labo); the random-effects
parameters and the residual variance (or scale parameter)
are probably unidentifiable
```

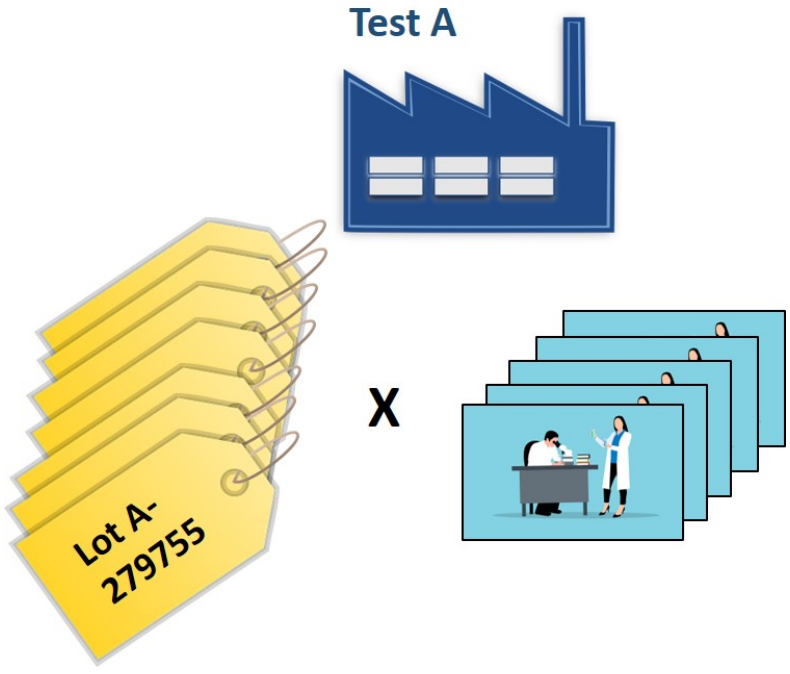
Modèle croisé aléatoire à 2 facteurs

Exemple

Pour quantifier la variabilité lié au lot de fabrication, on travaille sur un plus grand nombre de lots (10 lots), et on considère le facteur lot comme aléatoire. On demande à chaque laboratoire de faire 3 mesures sur chacun des 10 lots. On envoie à chacun des 15 laboratoires des kits issus des 10 lots du test, et 30 échantillons d'un même serum (donc 450 échantillons au total sont répartis aléatoirement entre les 15 laboratoires, et chaque laboratoire répartit aléatoirement les 30 échantillons reçus sur les 10 lots du test).



Facteurs



Deux facteurs croisés à prendre en compte :

- **laboratoire (aléatoire)**
- **lot (aléatoire)**

Modèle théorique

Deux modèles aléatoires peuvent en théorie être utilisés dans un tel cas, un modèle complet très rarement utilisé et un modèle simplifié négligeant l'interaction entre les deux facteurs aléatoires A et B.

- Modèle complet (très rarement utilisé):
  $X_{ijk} = \mu + A_i + B_j + AB_{ij} + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$,
  $A_i \sim N(0, \sigma_A)$, $B_j \sim N(0, \sigma_B)$ et $AB_{ij} \sim N(0, \sigma_{AB})$
- Modèle simplifié sans interaction (courant) :
  $X_{ijk} = \mu + A_i + B_j + \epsilon_{ijk}$ avec $\epsilon_{ijk} \sim N(0, \sigma)$, $A_i \sim N(0, \sigma_A)$ et
  $B_j \sim N(0, \sigma_B)$

Visualisation des données

```
str(d_c)
```

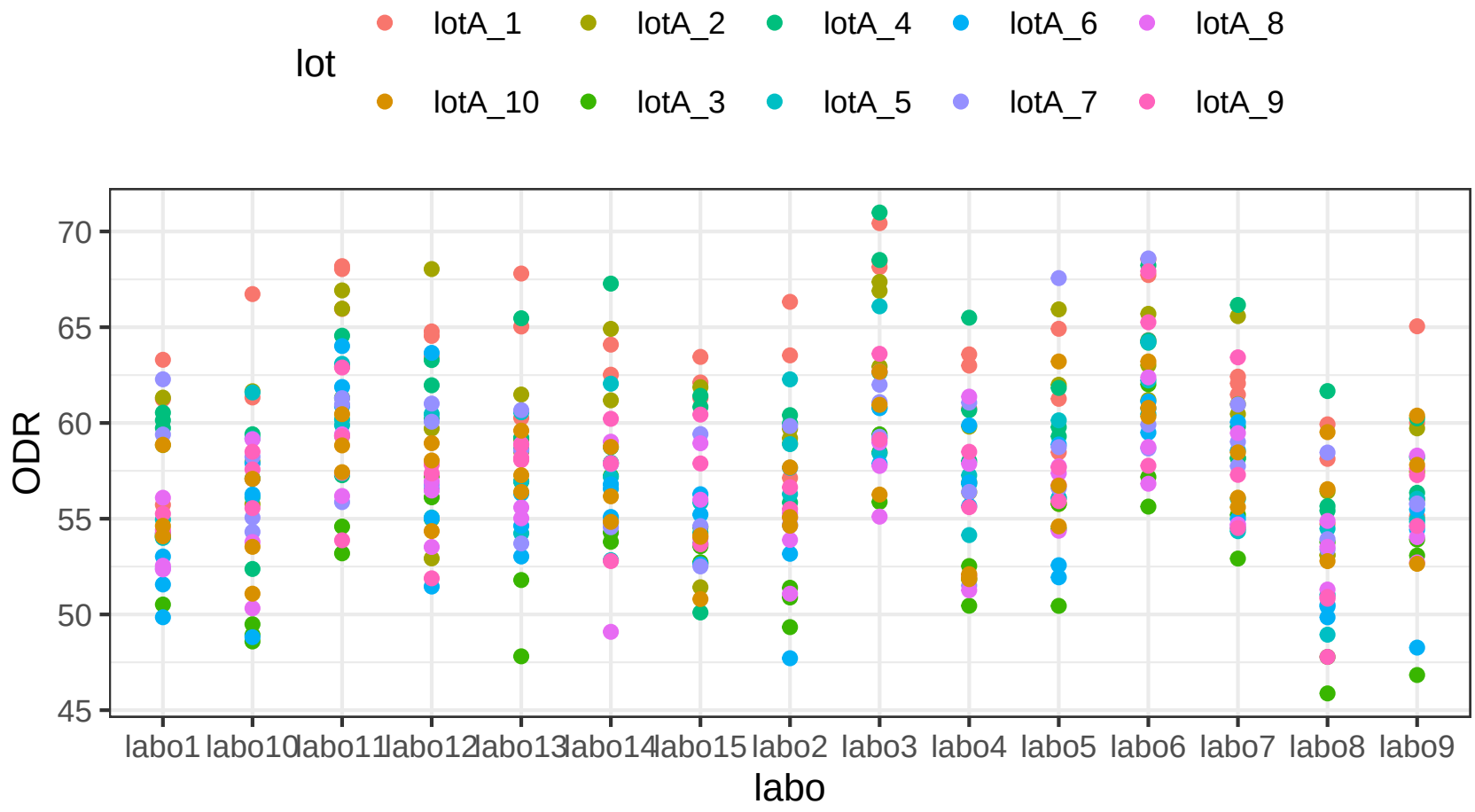
```
## 'data.frame':    450 obs. of  3 variables:
##  $ labo: Factor w/ 15 levels "labo1","labo10",..: 1 1 1 1 1 1
##  $ lot : Factor w/ 10 levels "lotA_1","lotA_10",..: 1 1 1 3 3
##  $ ODR : num  55.7 63.3 61.2 54.1 61.3 ...
```

```
xtabs(~ labo + lot, data = d_c)
```

```
##          lot
## labo     lotA_1 lotA_10 lotA_2 lotA_3 lotA_4 lotA_5 lotA_6 lo
##   labo1       3       3      3      3      3      3       3
##   labo10      3       3      3      3      3      3       3
##   labo11      3       3      3      3      3      3       3
##   labo12      3       3      3      3      3      3       3
##   labo13      3       3      3      3      3      3       3
##   labo14      3       3      3      3      3      3       3
##   labo15      3       3      3      3      3      3       3
##   labo2       3       3      3      3      3      3       3
##   labo3       3       3      3      3      3      3       3
```

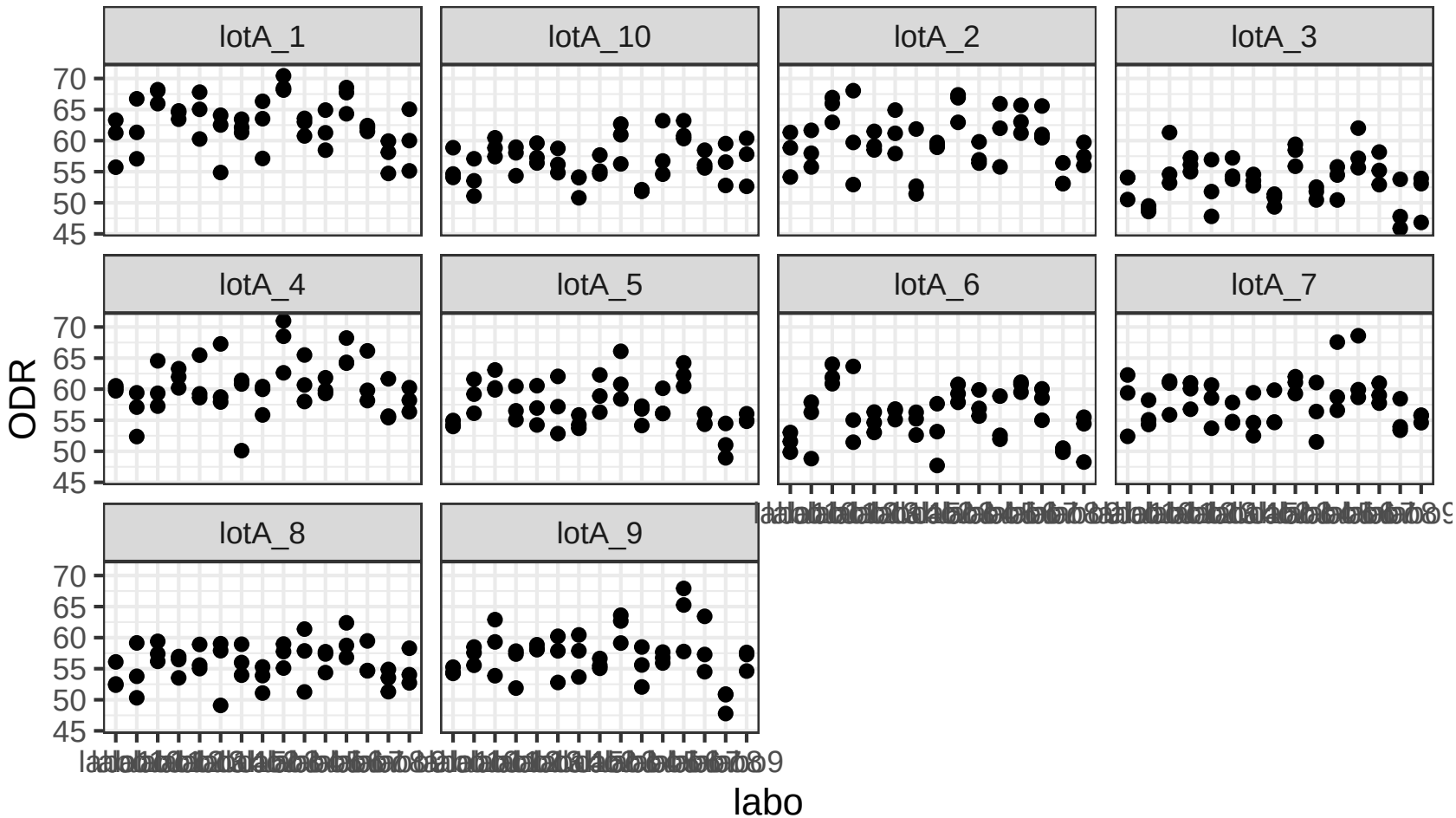
Représentations graphiques des données - une possible

```
ggplot(data = d_c, aes(x = labo, col = lot, y = ODR)) +
  geom_point() + theme(legend.position = "top")
```



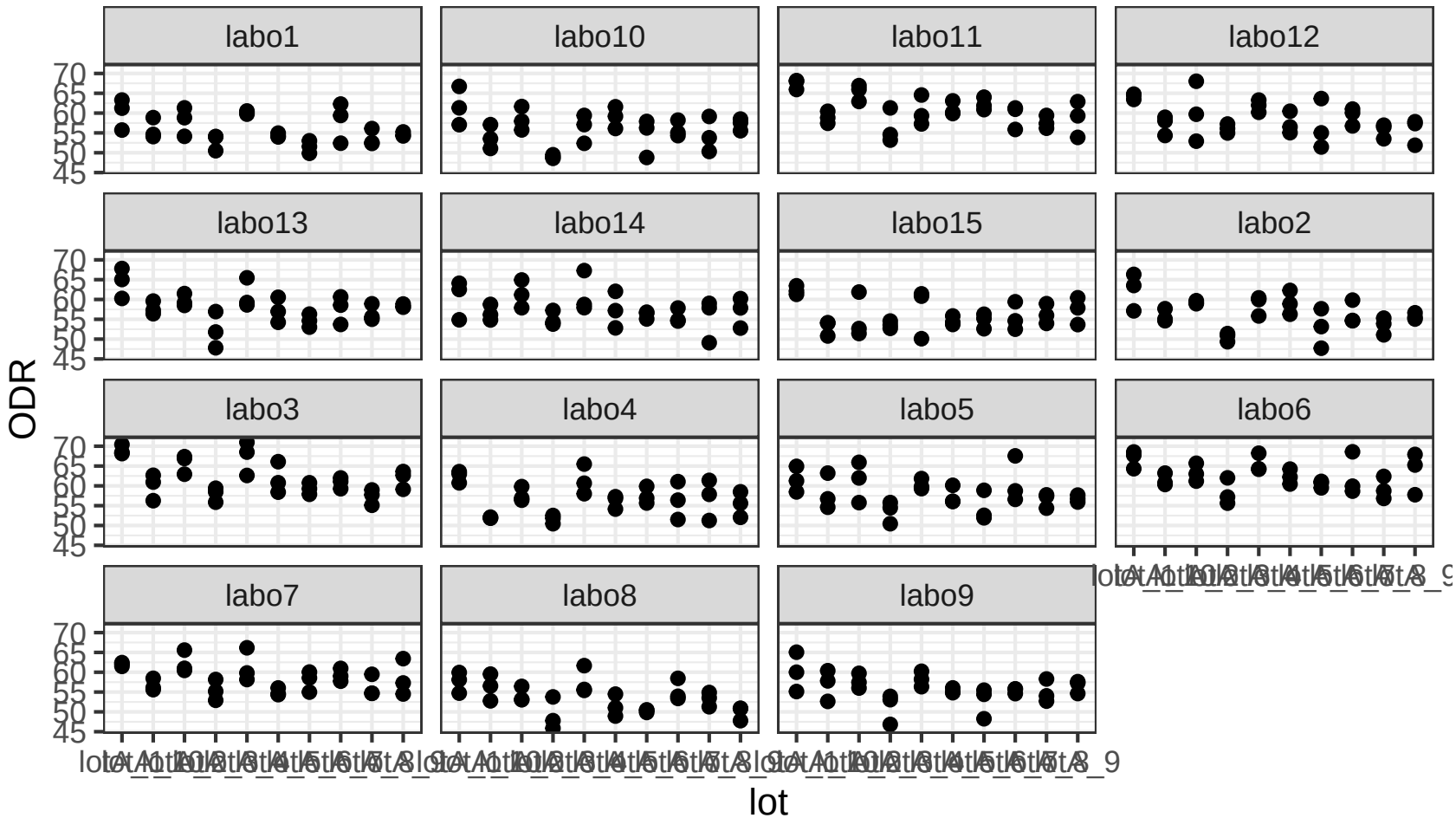
Représentations graphiques des données - une seconde

```
ggplot(data = d_c, aes(x = labo, y = ODR)) + geom_point() +
    facet_wrap(~ lot)
```



Représentations graphiques des données - une troisième

```
ggplot(data = d_c, aes(x = lot, y = ODR)) + geom_point() +
    facet_wrap(~ labo)
```



Ajustement du modèle aux données

```
(mm_c <- lmer(ODR ~ (1|lot) + (1|labo), data = d_c))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ (1 | lot) + (1 | labo)
##    Data: d_c
## REML criterion at convergence: 2339
## Random effects:
##  Groups   Name        Std.Dev.
##  labo     (Intercept) 2.39
##  lot      (Intercept) 2.68
##  Residual             2.99
## Number of obs: 450, groups:  labo, 15; lot, 10
## Fixed Effects:
## (Intercept)
##        57.8
```

Estimation des effets aléatoires : $\hat{\sigma} = 2.99$, $\hat{\sigma}_{labo} = 2.39$ et $\hat{\sigma}_{lot} = 2.68$

Intervalles de confiance

```
confint(mm_c)
```

```
##                2.5 % 97.5 %
## .sig01          1.66   3.60
## .sig02          1.73   4.37
## .sigma          2.80   3.21
## (Intercept) 55.62  59.88
```

Celui appelé `sigma` correspond toujours à l'écart type résiduel $\hat{\sigma}$ et les autres sont numérotés par ordre d'apparition dans le `print()` du modèle, donc ici `sigma01` correspond à $\hat{\sigma}_{labo}$ et `sigma02` à $\hat{\sigma}_{lot}$.

Graphe des résidus

```
plot(mm_c)
```

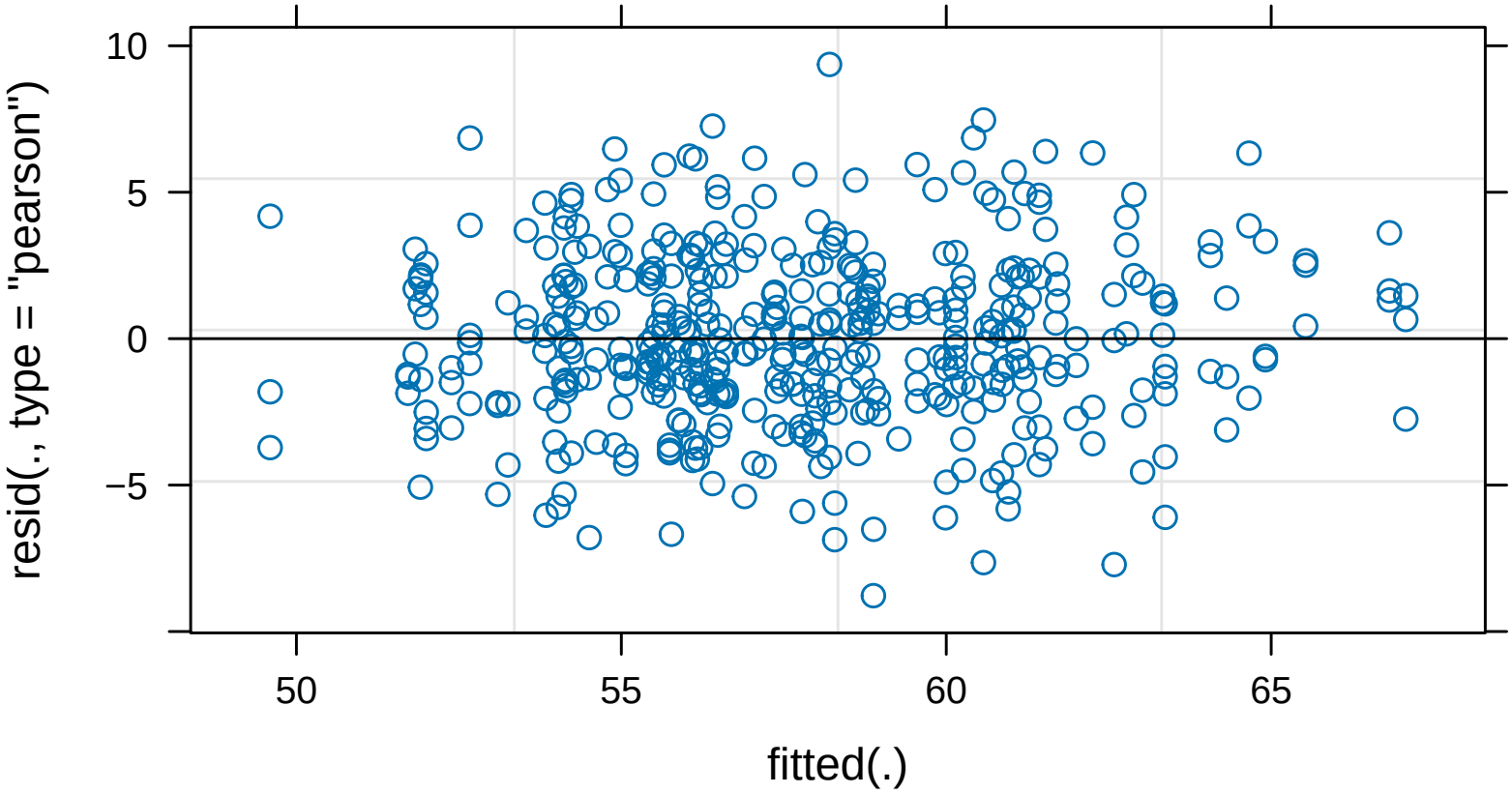


Diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm_c))
```

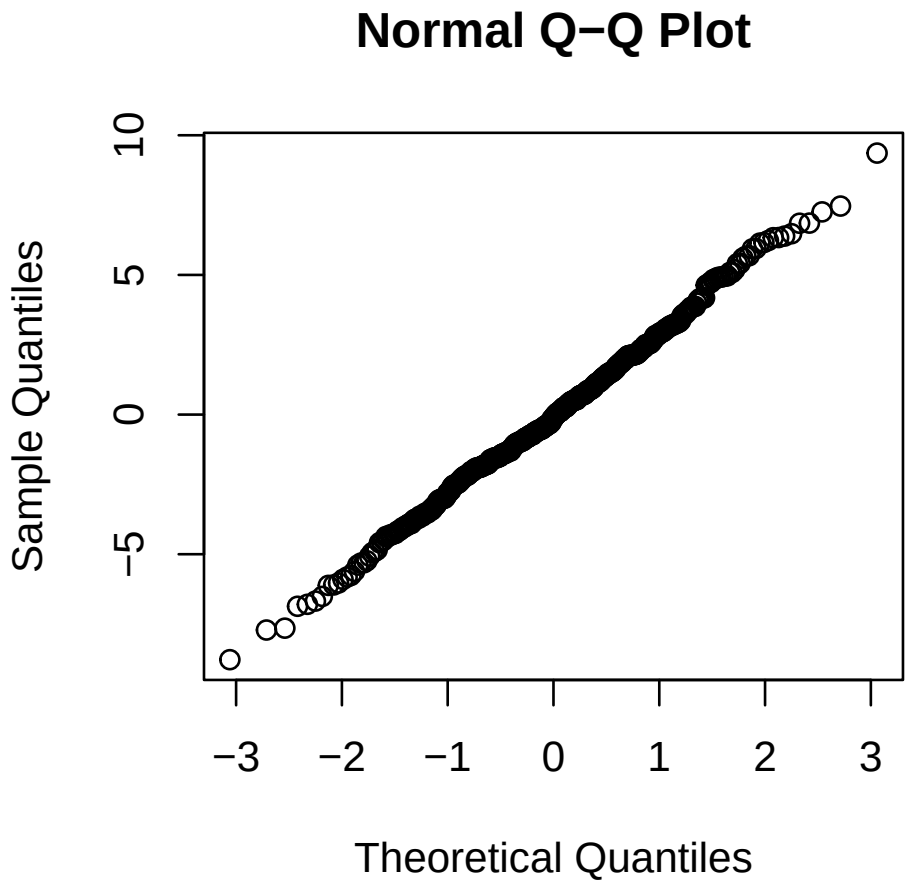


Diagramme quantile-quantile des effets aléatoires du labo

```
qqnorm(ranef(mm_c)$labo[, 1])
```

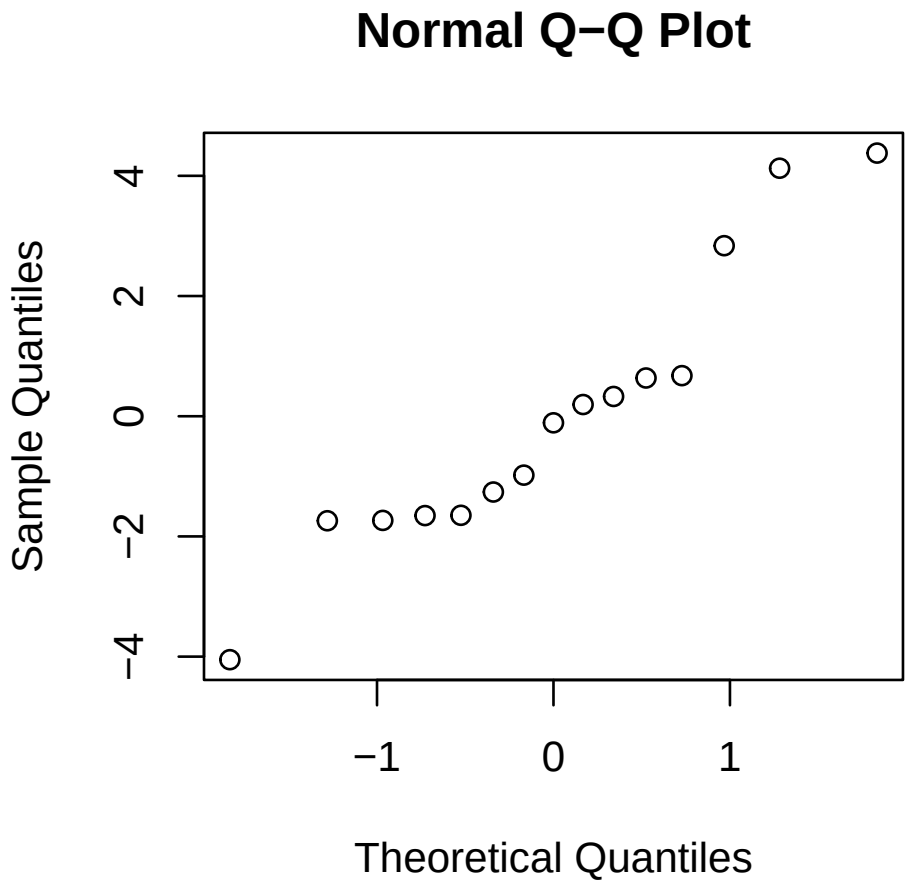
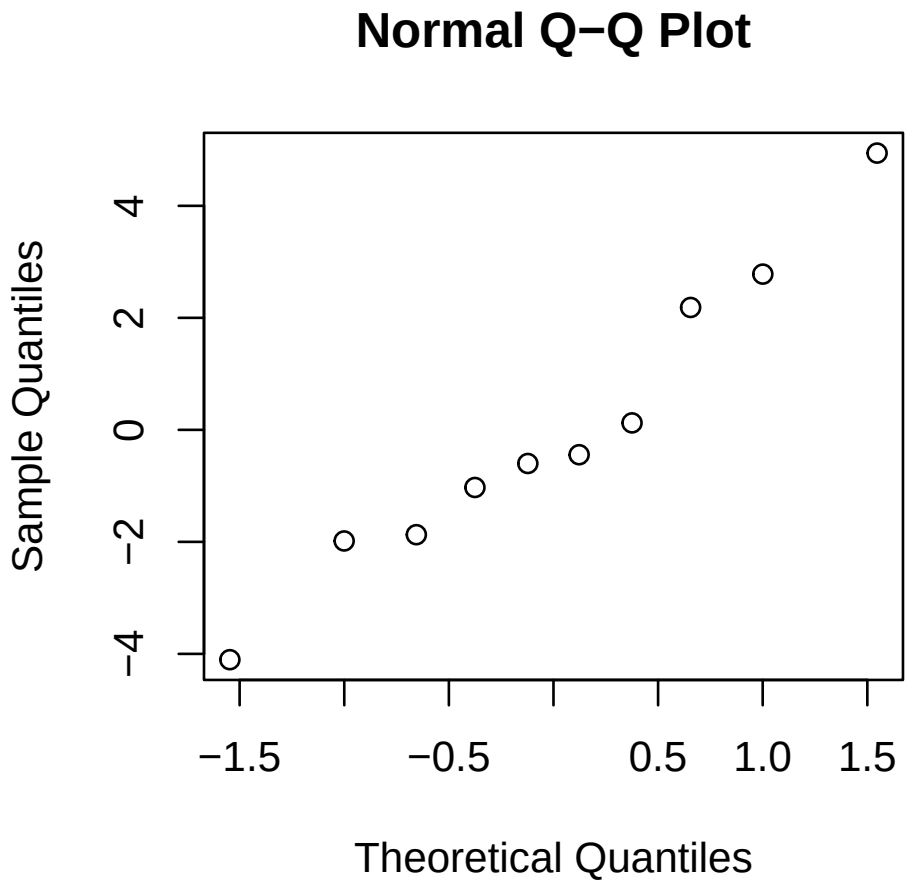


Diagramme quantile-quantile des effets aléatoires du lot

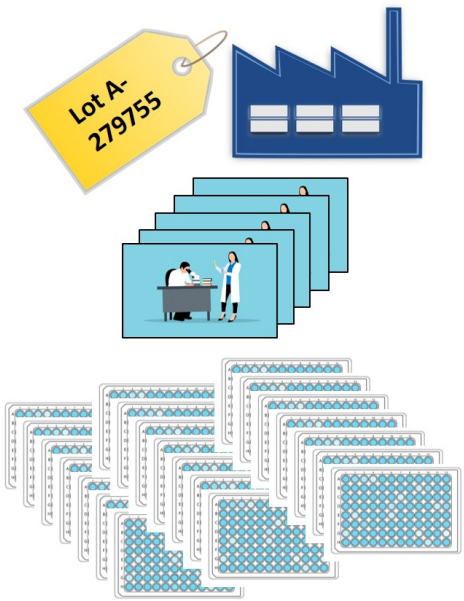
```
qqnorm(ranef(mm_c)$lot[, 1])
```



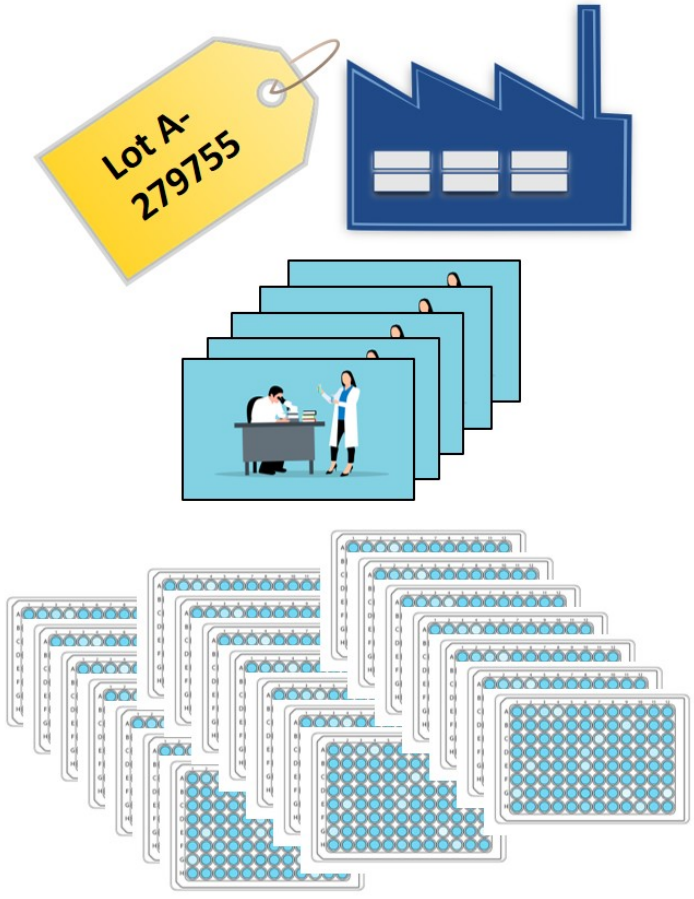
Modèle hiérarchisé aléatoire à 2 facteurs

Exemple

Dans un essai interlaboratoire, on souhaiterait prendre en compte l'effet plaque. Dans ce but on demande à chaque laboratoire de répartir les 30 échantillons d'un même serum qu'on lui envoie sur 10 plaques. Donc 450 échantillons au total sont répartis aléatoirement entre 15 laboratoires, et chaque laboratoire répartit aléatoirement les 30 échantillons reçus sur 10 plaques.



Facteurs



Deux facteurs imbriqués à prendre en compte :

- **laboratoire (aléatoire)**
- **plaque (aléatoire), imbriqué dans le laboratoire**

Modèle théorique

Un seul modèle peut être utilisé dans un tel cas, avec A aléatoire et B aléatoire imbriqué dans A et noté B(A).

$X_{ijk} = \mu + A_i + B(A)_{ij} + \epsilon_{ijk}$ avec
$\epsilon_{ijk} \sim N(0, \sigma)$,
$A_i \sim N(0, \sigma_A)$ et
$B(A)_{ij} \sim N(0, \sigma_{B(A)})$

Visualisation des données

```
str(d_d)
```

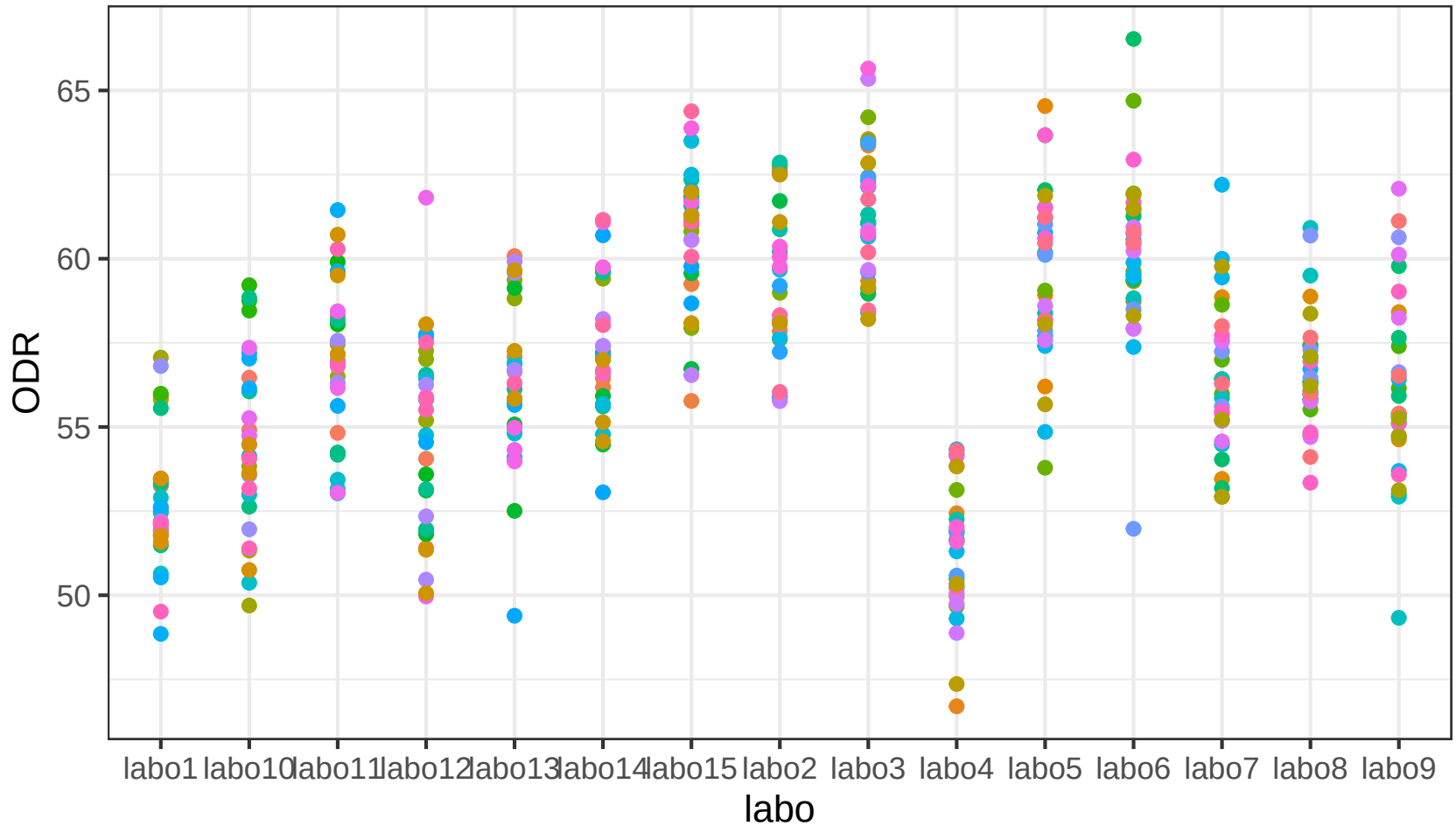
```
## 'data.frame':    450 obs. of  3 variables:
##  $ labo  : Factor w/ 15 levels "labo1","labo10",..: 1 1 1 1 1
##  $ plaque: Factor w/ 150 levels "P1_l1","P1_l10",..: 1 1 1 31
##  $ ODR   : num  53.2 52.4 52.6 57.1 55.8 ...
```

```
xtabs(~ plaque + labo, data = d_d)
```

```
##          labo
## plaque    labo1 labo10 labo11 labo12 labo13 labo14 labo15 lab
##   P1_l1       3      0      0      0      0      0      0
##   P1_l10      0      3      0      0      0      0      0
##   P1_l11      0      0      3      0      0      0      0
##   P1_l12      0      0      0      3      0      0      0
##   P1_l13      0      0      0      0      3      0      0
##   P1_l14      0      0      0      0      0      3      0
##   P1_l15      0      0      0      0      0      0      3
##   P1_l2       0      0      0      0      0      0      0
##   P1_l3       0      0      0      0      0      0      0
```

Représentation graphique des données

```
ggplot(data = d_d, aes(x = labo, col = plaque, y = ODR)) +
  geom_point() + theme(legend.position = "none")
```



Ajustement du modèle aux données

```
(mm_d <- lmer(ODR ~ (1|labo/plaque), data = d_d))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ (1 | labo/plaque)
##    Data: d_d
## REML criterion at convergence: 2070
## Random effects:
##  Groups       Name        Std.Dev.
##  plaque:labo (Intercept) 1.18
##  labo        (Intercept) 2.84
##  Residual                2.03
## Number of obs: 450, groups:  plaque:labo, 150; labo, 15
## Fixed Effects:
## (Intercept)
##          57
```

Estimation des effets aléatoires : $\hat{\sigma} = 2.03$, $\hat{\sigma}_{labo} = 2.84$ et $\hat{\sigma}_{plaque} = 1.18$

Intervalles de confiance

```
confint(mm_d)
```

```
##                 2.5 % 97.5 %
## .sig01          0.881   1.48
## .sig02          1.960   4.16
## .sigma          1.881   2.21
## (Intercept) 55.441  58.46
```

Celui appelé `sigma` correspond toujours à l'écart type résiduel $\hat{\sigma}$ et les autres sont numérotés par ordre d'apparition dans le `print()` du modèle, donc ici `sigma01` correspond à $\hat{\sigma}_{plaque}$ et `sigma02` à $\hat{\sigma}_{labo}$.

Graphe des résidus

```
plot(mm_d)
```

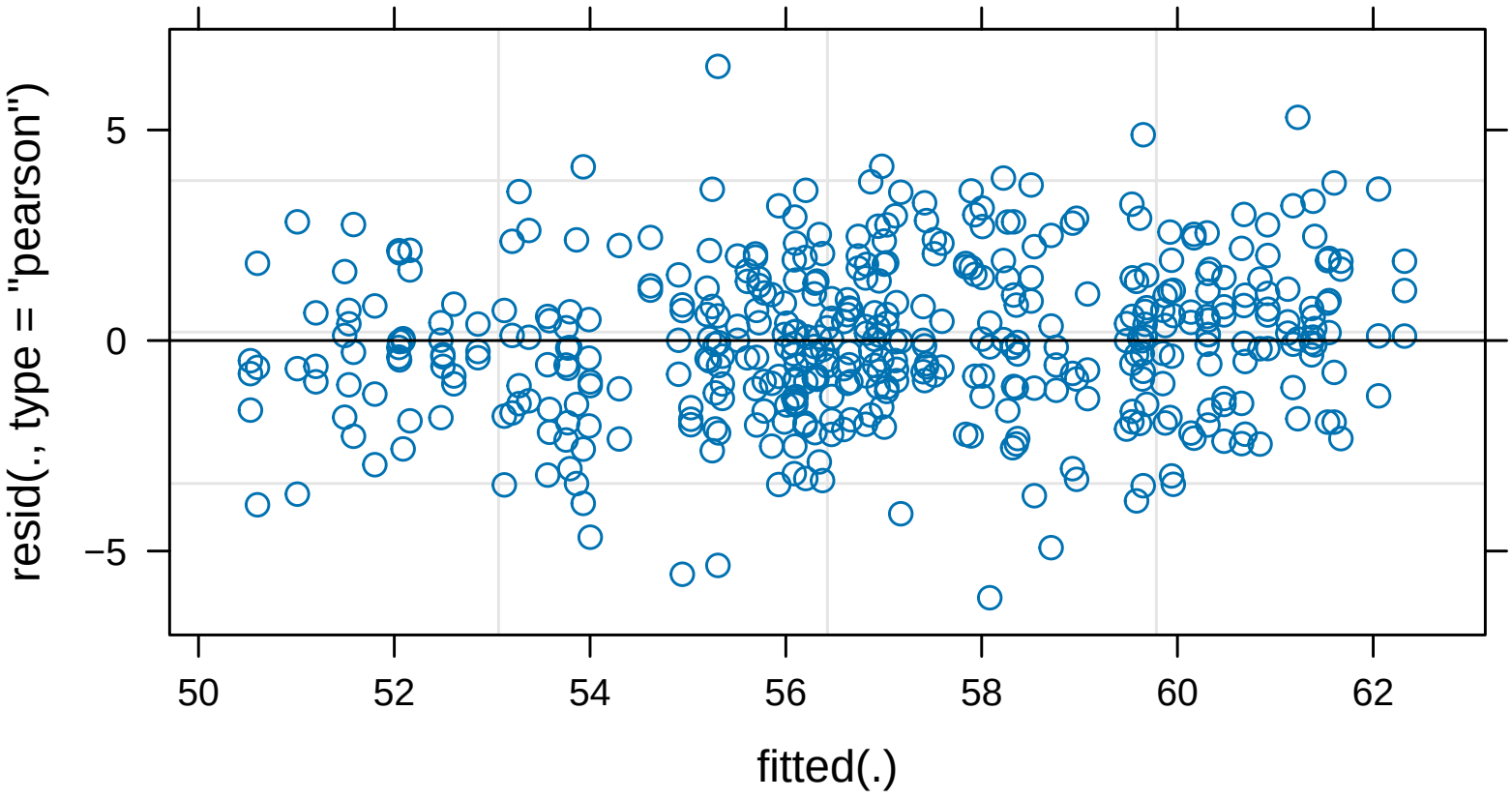


Diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm_d))
```

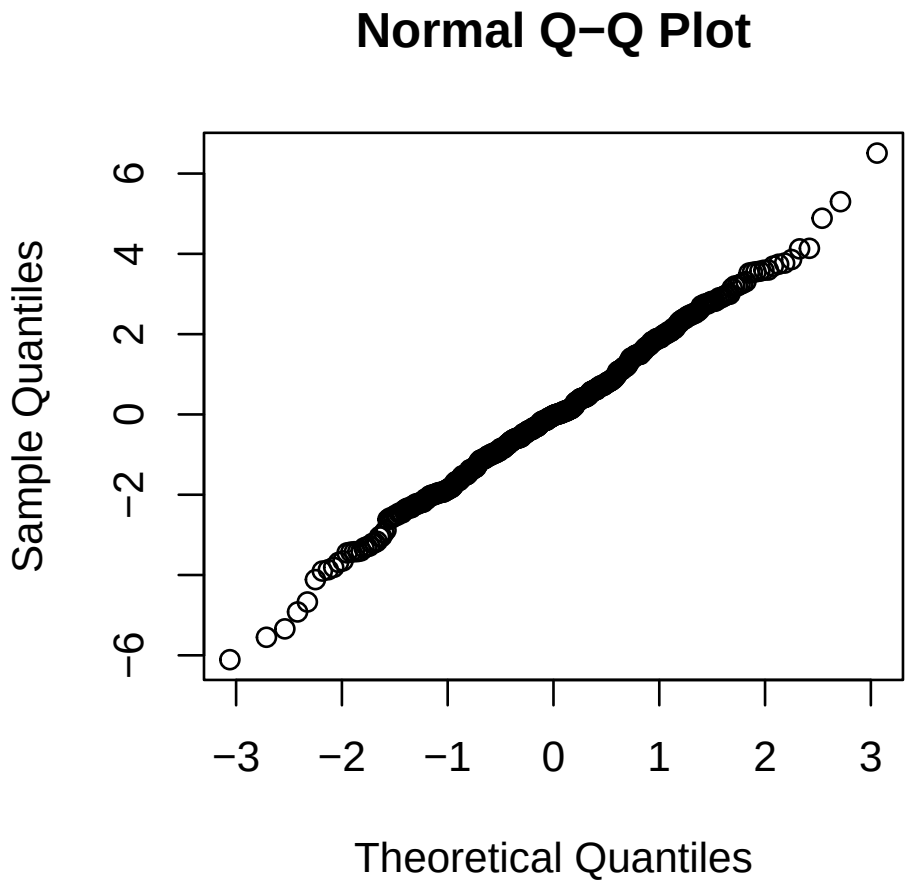


Diagramme quantile-quantile des effets aléatoires du labo

```
qqnorm(ranef(mm_d)$labo[, 1])
```

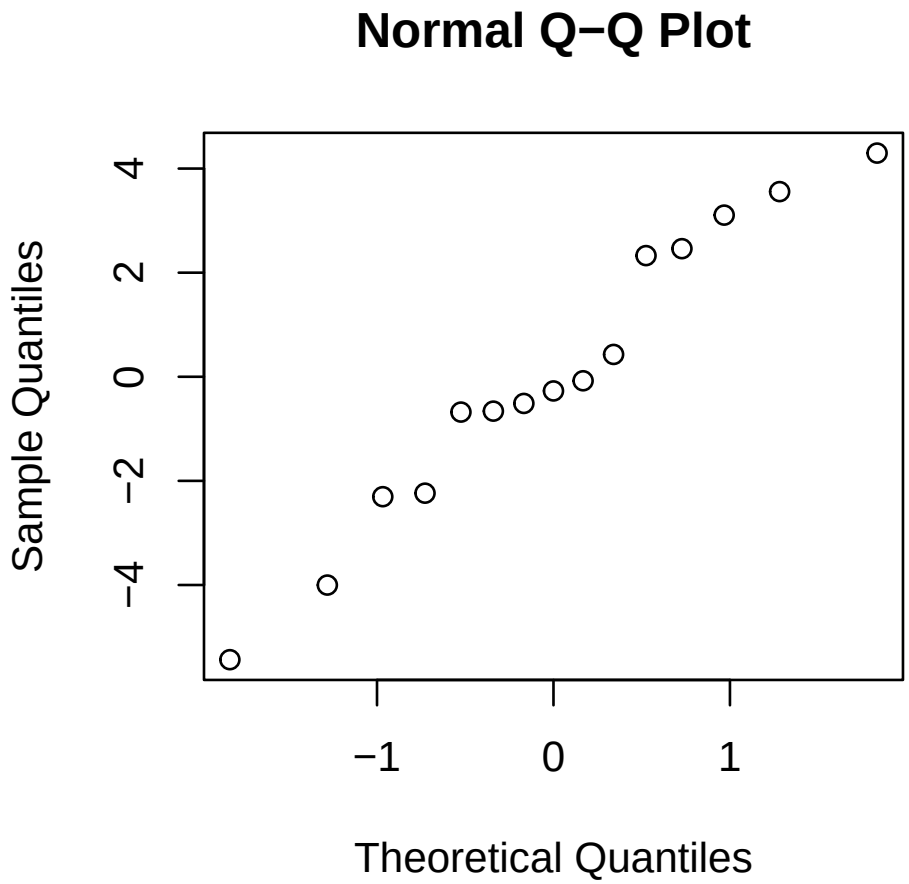
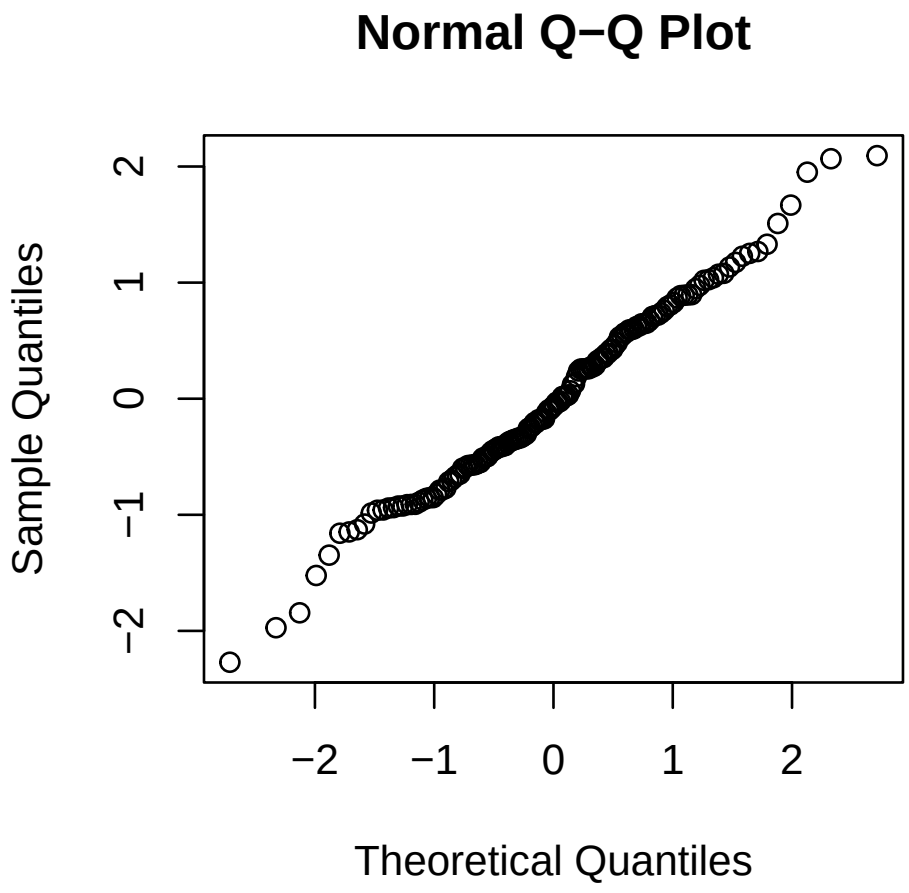


Diagramme quantile-quantile des effets aléatoires de la plaque

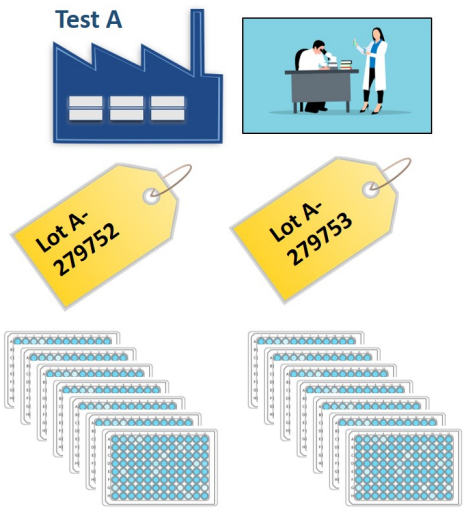
```
qqnorm(ranef(mm_d)$plaque[, 1])
```



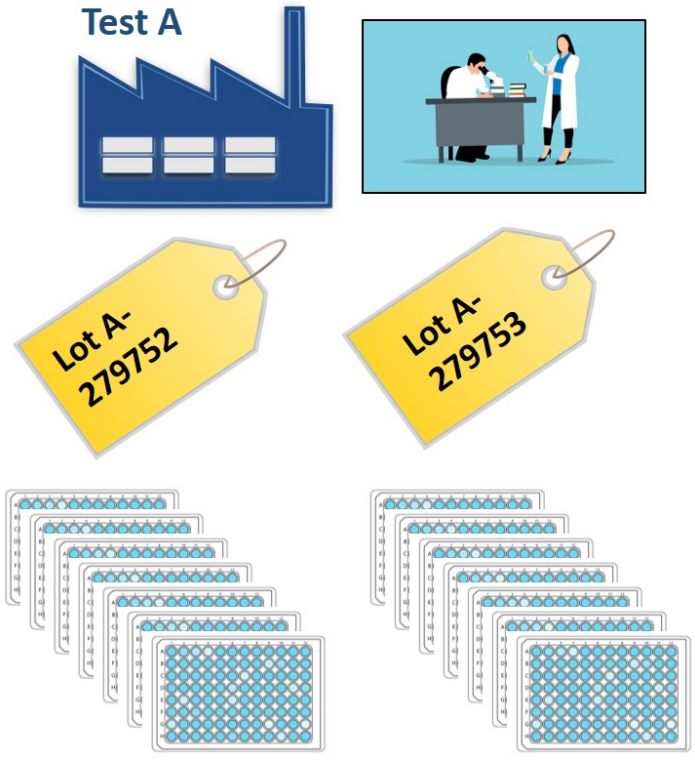
Modèle hiérarchisé mixte à 2 facteurs

Exemple

Un laboratoire souhaite contrôler la stabilité des lots qui lui sont vendus par son fabricant. A un instant t il ne dispose généralement que d'un seul lot. Imaginons qu'à chaque réception d'un nouveau lot, il fasse 50 mesures sur 50 échantillons d'un même serum de contrôle (qui ne bouge pas dans le temps), sur 5 plaques différentes (pour prendre en compte l'effet plaque), et qu'il les compare aux mêmes mesures qu'il avait faites sur le même serum avec le lot précédent.



Facteurs



Deux facteurs imbriqués à prendre en compte :

- **lot (ici fixe, on ne veut comparer que deux lots entre eux)**
- **plaque (aléatoire), imbriqué dans le lot**

Modèle théorique

Un seul modèle être utilisé dans un tel cas, avec A fixe et B aléatoire imbriqué dans A et noté B(A).

$X_{ijk} = \mu + \alpha_i + B(A)_{ij} + \epsilon_{ijk}$
avec
$\epsilon_{ijk} \sim N(0, \sigma)$ et
$B(A)_{ij} \sim N(0, \sigma_{B(A)})$

Visualisation des données

```
str(d_e)
```

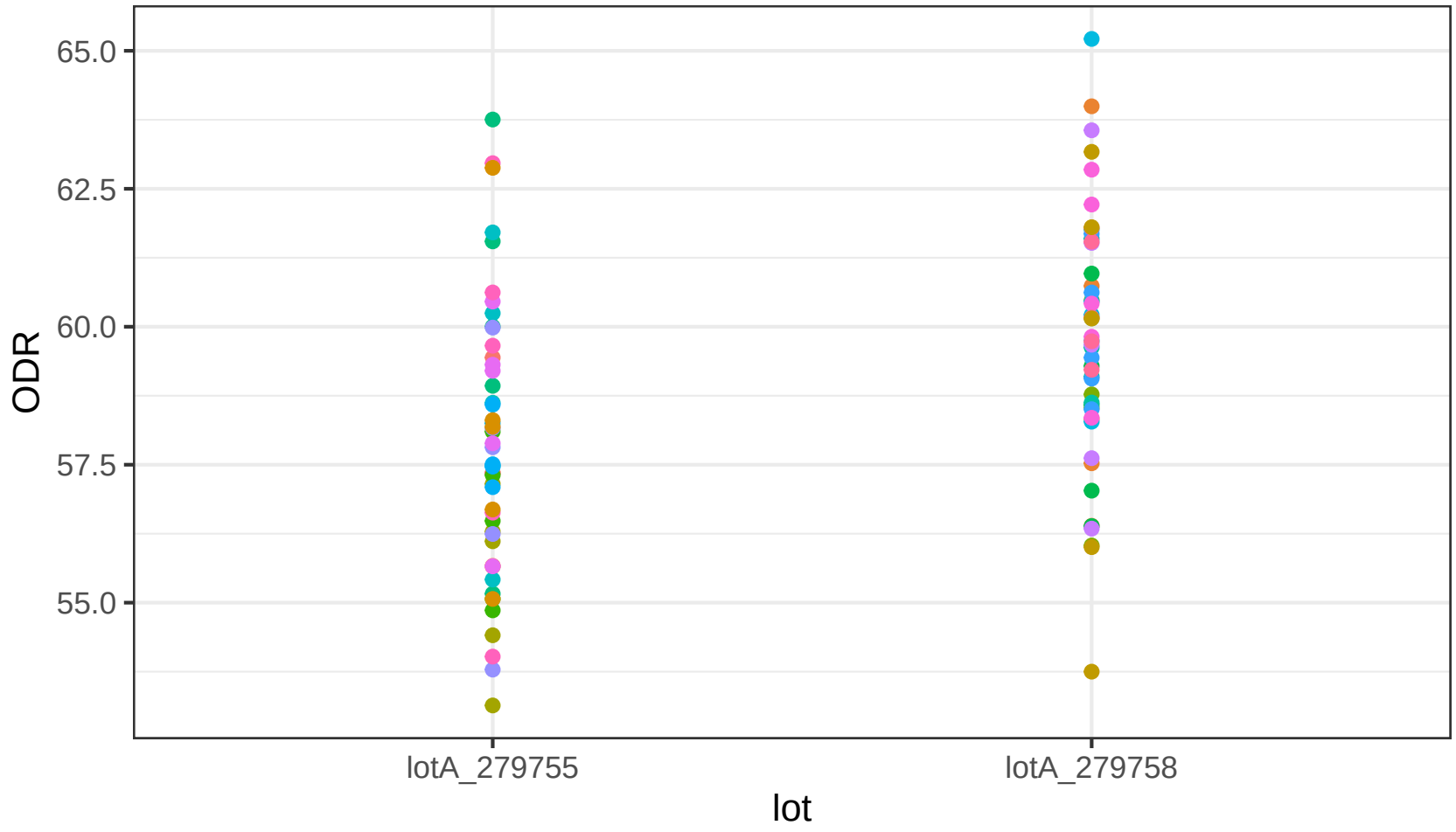
```
## 'data.frame':    100 obs. of  3 variables:
##  $ labo  : Factor w/ 2 levels "lotA_279755",..: 1 2 1 2 1 2 1
##  $ plaque: Factor w/ 20 levels "P1_lotA_279755",..: 1 2 1 2 1
##  $ ODR   : num  57.3 56.4 59.4 58.3 55.7 ...
```

```
xtabs(~ plaque + lot, data = d_e)
```

```
##                   lot
## plaque             lotA_279755 lotA_279758
##   P1_lotA_279755              5           0
##   P1_lotA_279758              0           5
##   P10_lotA_279755             5           0
##   P10_lotA_279758             0           5
##   P2_lotA_279755              5           0
##   P2_lotA_279758              0           5
##   P3_lotA_279755              5           0
##   P3_lotA_279758              0           5
##   P4_lotA_279755              5           0
```

Une représentation graphique des données

```
ggplot(data = d_e, aes(x = lot, col = plaque, y = ODR)) +
  geom_point() + theme(legend.position = "none")
```



Ajustement du modèle aux données

```
(mm_e <- lmer(ODR ~ lot + (1|plaque), data = d_e))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ODR ~ lot + (1 | plaque)
##    Data: d_e
## REML criterion at convergence: 453
## Random effects:
##  Groups   Name        Std.Dev.
##  plaque   (Intercept) 0.33
##  Residual             2.32
## Number of obs: 100, groups:  plaque, 20
## Fixed Effects:
##    (Intercept)  lotlotA_279758
##          57.79            1.94
```

Estimation des effets fixes : $\hat{\mu}_{lotA_{279755}} = 57.79$ et $\hat{\mu}_{lotA_{279758}} = 57.79 + 1.94$

Estimation des effets aléatoires : $\hat{\sigma} = 2.32$, $\hat{\sigma}_{plaque} = 0.33$

Intervalles de confiance

```
confint(mm_e)
```

```
##                   2.5 % 97.5 %
## .sig01             0.00   1.04
## .sigma             2.03   2.69
## (Intercept)       57.14  58.44
## lotlotA_279758     1.02   2.86
```

Celui appelé `sigma` correspond toujours à l'écart type résiduel $\hat{\sigma}$ donc ici `sigma01` correspond à $\hat{\sigma}_{plaque}$.

Graphe des résidus

```
plot(mm_e)
```

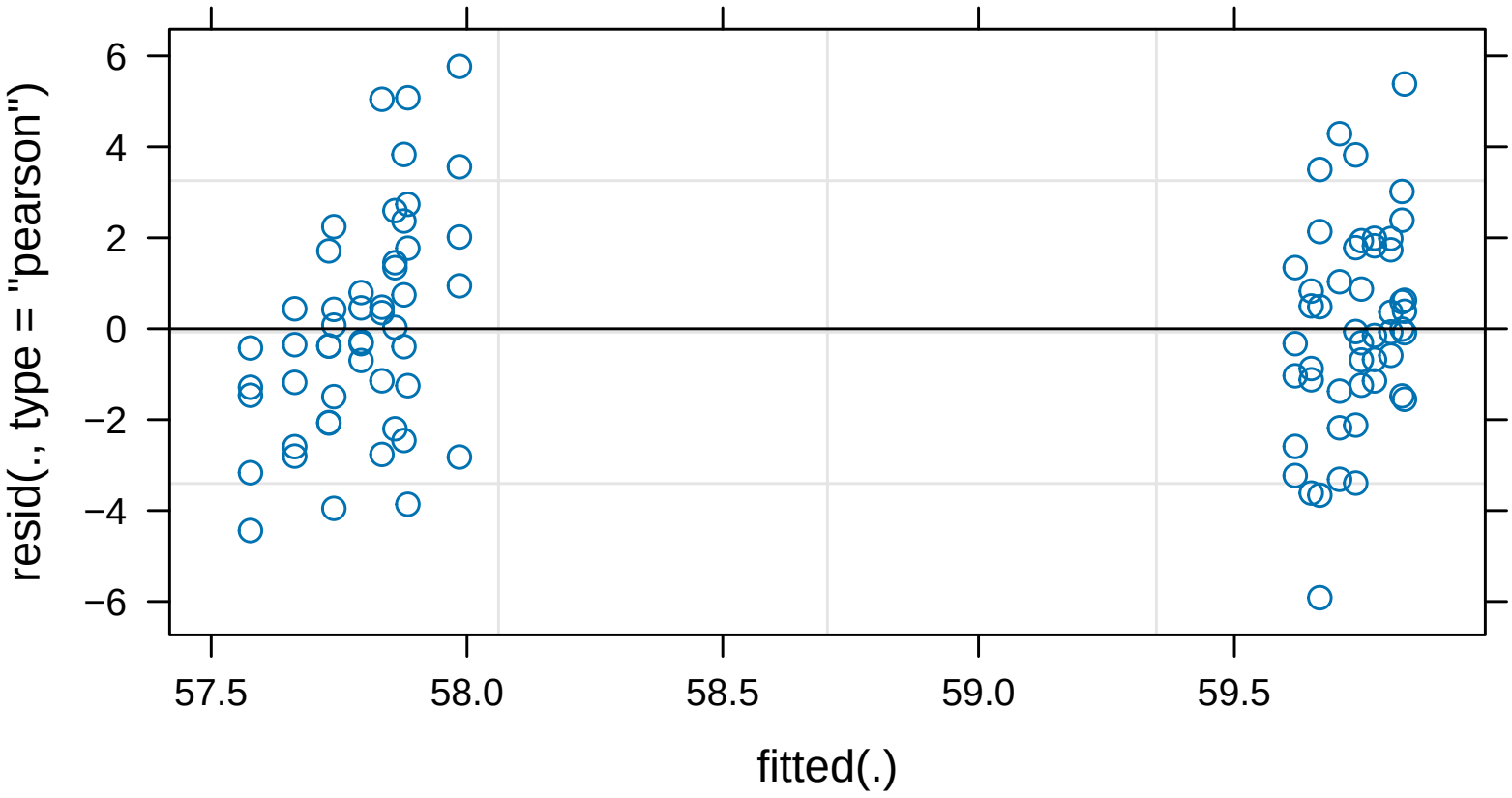


Diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm_e))
```

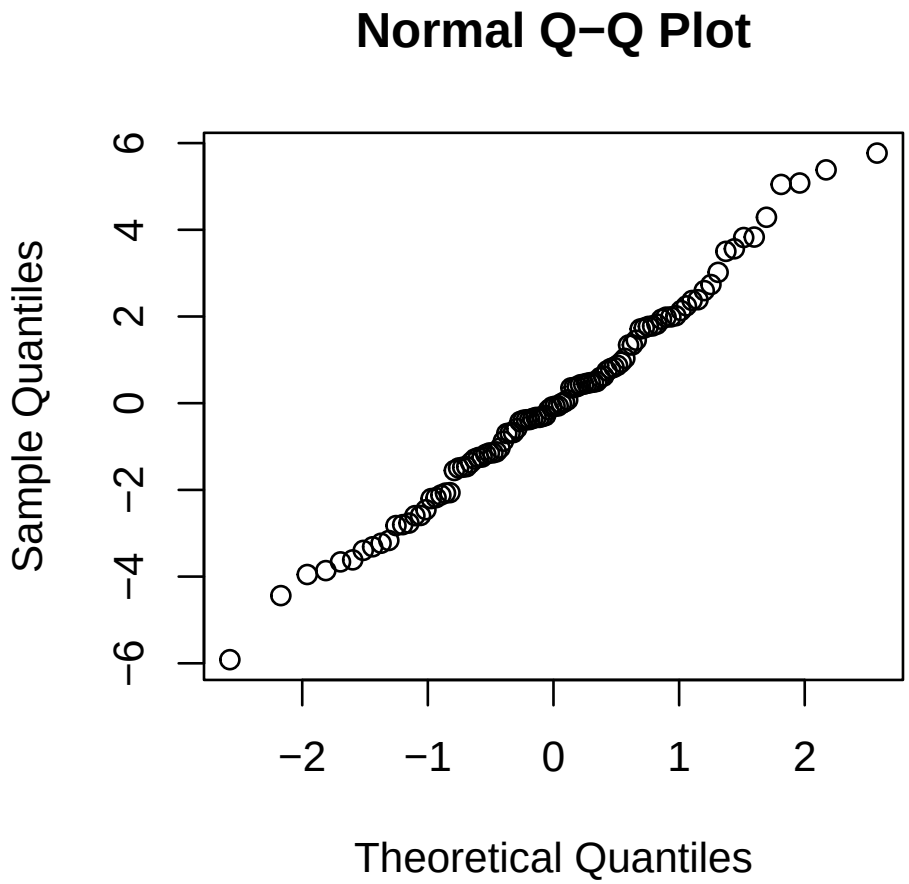
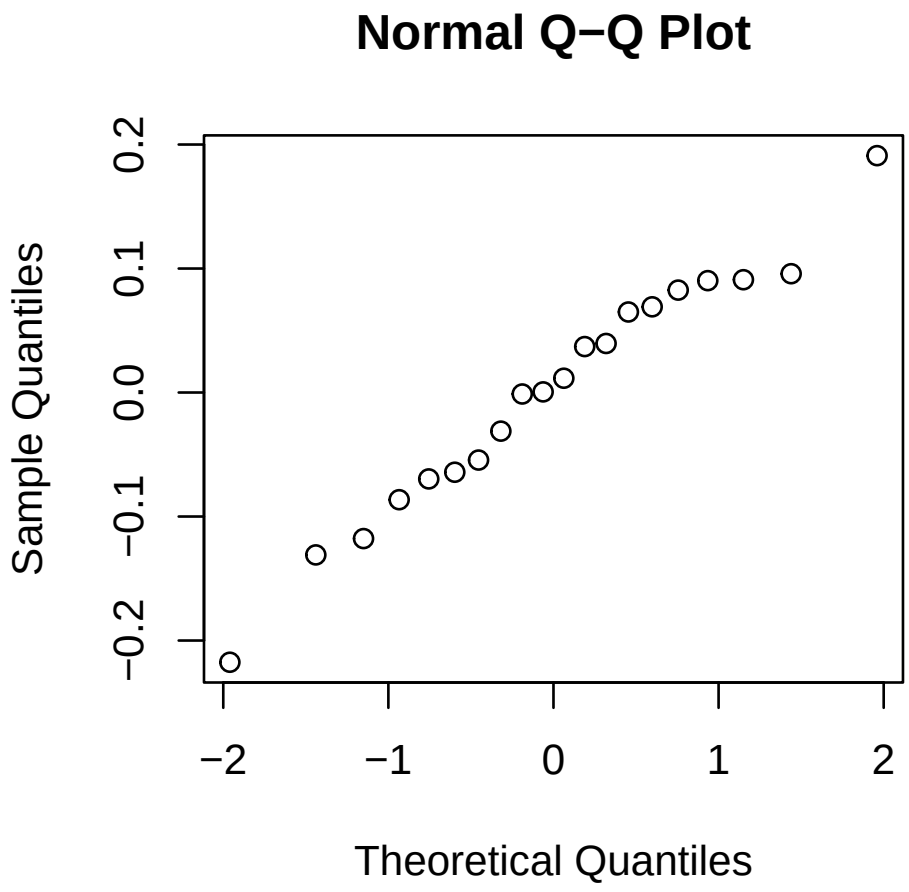


Diagramme quantile-quantile des effets aléatoires du plaque

```
qqnorm(ranef(mm_e)$plaque[, 1])
```



Points divers

Quelques remarques au sujet des modèles hiérarchisés

- Erreur fréquente : utiliser les mêmes numéros de plaque d'un labo à l'autre ou d'un lot à l'autre
  - Bien regarder le design avec `xtabs()`
  - Fabriquer un nouveau facteur identifiant si besoin
- Que se passe-t-il si on n'a pas de répétition sur le facteur imbriqué (ex. une seule mesure par plaque) ?

Quelques remarques au sujet des modèles hiérarchisés

- Erreur fréquente : utiliser les mêmes numéros de plaque d'un labo à l'autre ou d'un lot à l'autre
  - Bien regarder le design avec `xtabs()`
  - Fabriquer un nouveau facteur identifiant si besoin
- Que se passe-t-il si on n'a pas de répétition sur le facteur imbriqué (ex. une seule mesure par plaque) ?
  On est ramené à un modèle d'ANOVA1.

Absence de valeurs de p dans les sorties de `lmer()`

Exemple :

```
summary(mm_e)$coefficients
```

```
##                  Estimate Std. Error t value
## (Intercept)         57.79      0.345  167.59
## lotlotA_279758       1.94      0.488    3.98
```

`lmer()` ne fournit pas de valeurs de p associées à chacun des paramètres du modèle (ddl de la loi de Student associée non connu - le package `lmerTest` peut être utilisé pour cela (si vous y tenez absolument) mais `lmer()` fournit des **intervalles de confiance** sur les paramètres du modèle (effet fixe ou aléatoire) qui sont plus faciles à calculer et **beaucoup plus informatifs**.

Tests sur les effets aléatoires ?

L'estimation des composantes de la variance est généralement plus intéressante que le test des effets aléatoires.
Néanmoins `lmer()` permet de tester un effet aléatoire par le biais de la comparaison de 2 modèles emboîtés à l'aide de la **statistique du rapport de vraisemblance** (la **partie fixe des 2 modèles** doit être **identique** et l'argument `REML` doit être fixé à `FALSE`).
Mais il serait maladroit de se baser sur le test d'un effet aléatoire pour savoir si on introduit cet effet dans le modèle ou non.
Un effet aléatoire qui a un sens biologique, même s'il est non significatif, doit être pris en compte dans l'estimation des effets fixes si c'est techniquement possible.

Récapitulatif des modèles à 2 facteurs

Formulation des différents modèles sous R

- A et B aléatoires croisés sans interaction : `(1|A) + (1|B)`
- A fixe et B aléatoire croisés avec interaction : `A + (A|B)`
- A fixe et B aléatoire croisés sans interaction : `A + (1|B)`
- A aléatoire et B aléatoire emboîté dans A : `(1|A/B)`
- A fixe et B aléatoire emboîté dans A : `A + (1|B)`

Les formules sont les mêmes pour une variable explicative quantitative dont l'effet modélisé est forcément un effet fixe (donc codé dans la formule comme un facteur fixe), mais **on suppose alors une relation linéaire et l'effet estimé est le coefficient de régression** (pente) de cette relation linéaire.

Modèles à plus de 2 facteurs

Modèles à plus de 2 facteurs

- Les concepts vus précédemment permettent en théorie de traiter tous les modèles à plus de 2 facteurs.
- Le nombre de modèles possibles avec plus de 2 facteurs devient vite grand et le choix d'un modèle plus compliqué (quelles interactions prendre en compte ?).
- l'interprétation des résultats se complique aussi avec l'augmentation du nombre de facteurs et/ou l'augmentation du nombre de modalités par facteur.

Il est généralement recommandé de limiter autant que possible le nombre de facteurs étudiés dans une expérience biologique et de demander un avis statistique dès la planification expérimentale lorsque le nombre de facteurs augmente.

Un exemple de modèle mixte à 3 facteurs

On mesure la **teneur en ASAT** dans le sang (aspartate aminotransférase) chez 64 vaches de deux races différentes, à quatre stades de lactation différents, PS0 (tarissement), PS1 (vêlage), PS2 (pic de lactation) et PS3 (milieu de lactation).

On souhaite étudier l'effet **race** et l'effet **stade de lactation** en prenant en compte bien entendu l'effet **vache** (chaque vache faisant l'objet de plusieurs mesures à différents stades de lactation).

Quel modèle ?

Un exemple de modèle mixte à 3 facteurs

On mesure la **teneur en ASAT** dans le sang (aspartate aminotransférase) chez 64 vaches de deux races différentes, à 4 stades de lactation différents, PS0 (tarissement), PS1 (vêlage), PS2 (pic de lactation) et PS3 (milieu de lactation).

On souhaite étudier l'effet **race** et l'effet **stade de lactation** en prenant en compte bien entendu l'effet **vache** (chaque vache faisant l'objet de plusieurs mesures à différents stades de lactation).

Modèle mixte avec **2 facteurs fixes croisés (stade et race)** et **un facteur aléatoire (vache)** qui est **imbriqué dans le facteur race** et **croisé avec le facteur stade**.

Visualisation des données

Données issues de la thèse vétérinaire d'Elodie Magnien : "Evaluation de l'activité de deux enzymes hépatiques (ASAT et GGT) au cours du cycle de production de la vache laitière", 2016, VetAgro Sup.

```
d <- read.table("DATA/ASAT.txt", header = TRUE,
                stringsAsFactors = TRUE)
str(d)
```

```
## 'data.frame':    238 obs. of  4 variables:
##  $ vache  : Factor w/ 64 levels "V1030","V1053",..: 27 30 51
##  $ stade  : Factor w/ 4 levels "PS0","PS1","PS2",..: 1 1 1 1
##  $ race   : Factor w/ 2 levels "MTB","PH": 1 1 1 1 1 1 1 1 1
##  $ LogASAT: num  1.97 2.08 2.08 1.99 2.39 ...
```

Plan d'expérience

```
xtabs(~ race + stade, data = d)
```

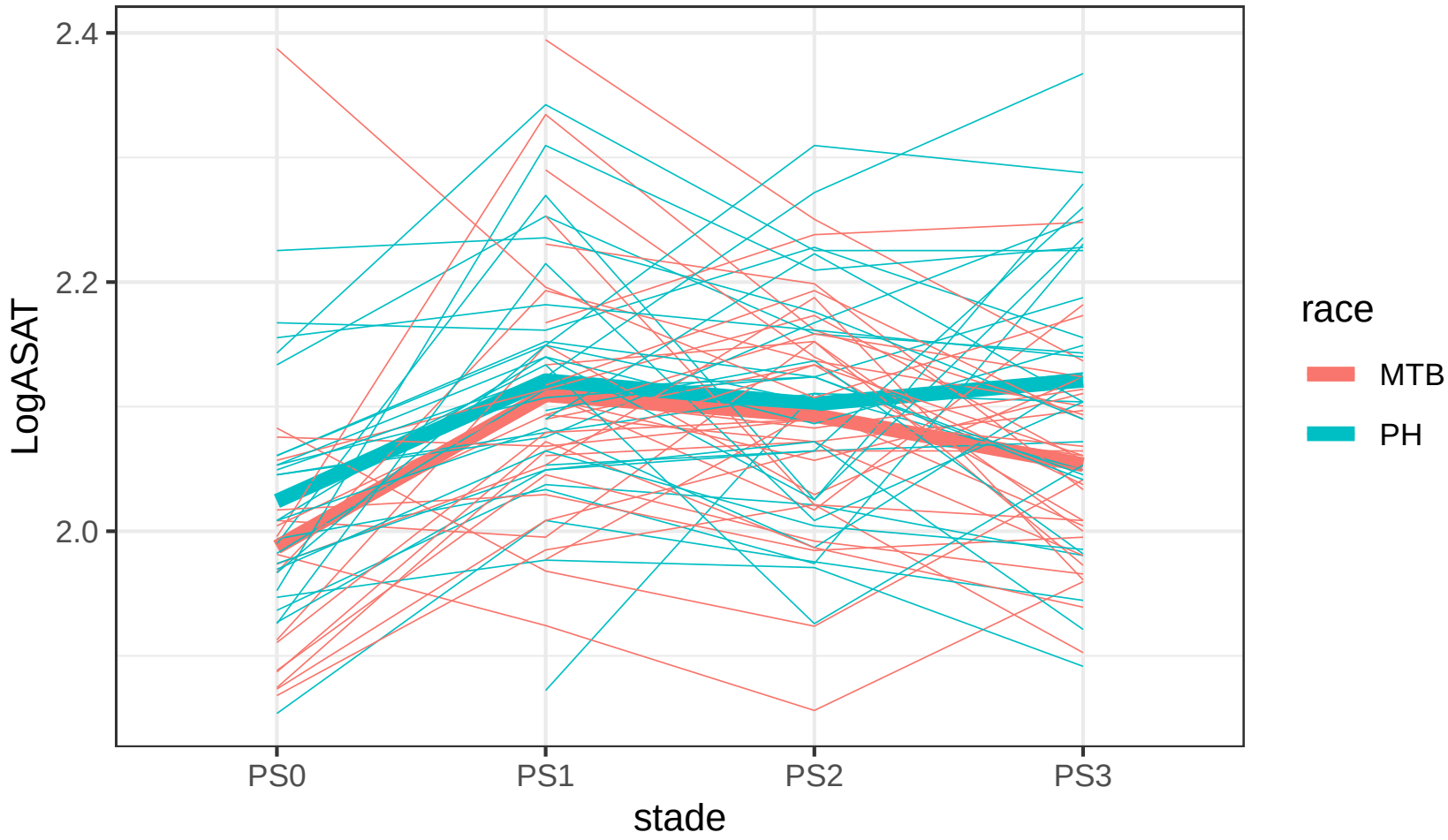
```
##       stade
## race   PS0 PS1 PS2 PS3
##   MTB   19  33  33  33
##   PH    27  31  31  31
```

```
xtabs(~ vache + stade, data = d)
```

```
##         stade
## vache    PS0 PS1 PS2 PS3
##   V1030    1   1   1   1
##   V1053    1   1   1   1
##   V1063    0   1   1   1
##   V1224    1   1   1   1
##   V130     1   1   1   1
##   V2007    0   1   1   1
##   V2328    1   1   1   1
##   V2329    0   1   1   1
##   V2331    1   1   1   1
##   V2333    1   1   1   1
```

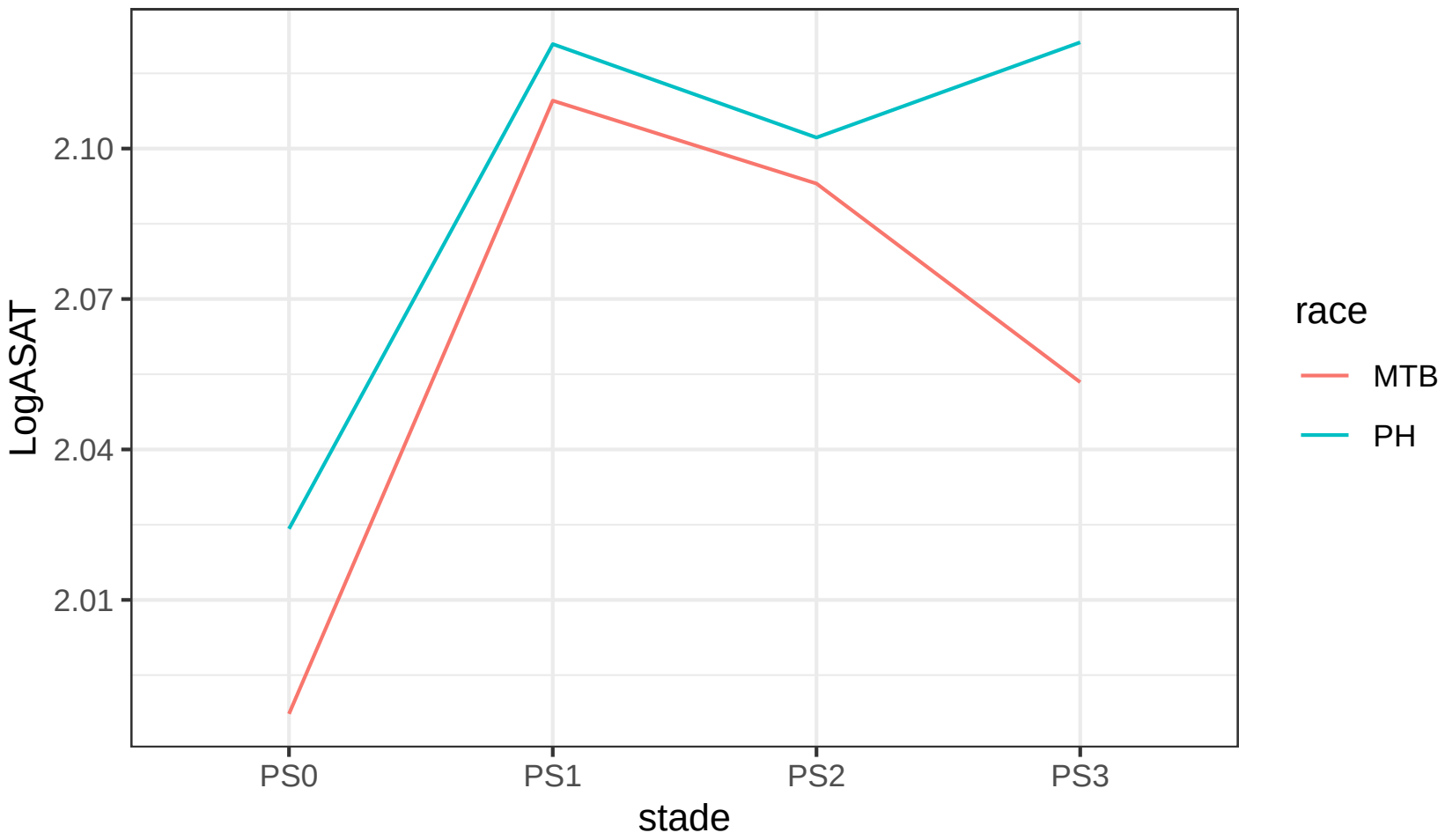
Tracés individuels

```
ggplot(data = d, aes(x = stade, y = LogASAT, colour = race,
  group = vache)) + stat_summary(aes(group = race),
  geom = "line", lwd = 2, fun = mean) + geom_line(lwd = 0.2)
```



Graphe d'interaction sans données individuelles

```
ggplot(data = d, aes(x = stade, y = LogASAT,
  colour = race)) + stat_summary(fun = mean, geom = "line",
                aes(group = race))
```



A votre tour

Ajustez un (ou des modèles) à ces données en vue de caractériser l'effet **race** et l'effet **stade** de lactation sur la **teneur en ASAT**.

Ajustement du modèle aux données

```
mm <- lmer(LogASAT ~ race + stade + race:stade + (1|vache),
           data = d)
summary(mm)$coefficients
```

```
##                   Estimate Std. Error t value
## (Intercept)        2.00094     0.0212  94.432
## racePH             0.02049     0.0282   0.728
## stadePS1           0.10862     0.0226   4.817
## stadePS2           0.09209     0.0226   4.083
## stadePS3           0.05247     0.0226   2.327
## racePH:stadePS1 -0.00921     0.0303  -0.304
## racePH:stadePS2 -0.01131     0.0303  -0.374
## racePH:stadePS3  0.04729     0.0303   1.563
```

```
summary(mm)$varcor
```

```
##  Groups   Name        Std.Dev.
##  vache    (Intercept) 0.0618
##  Residual             0.0761
```

Intervalles de confiance

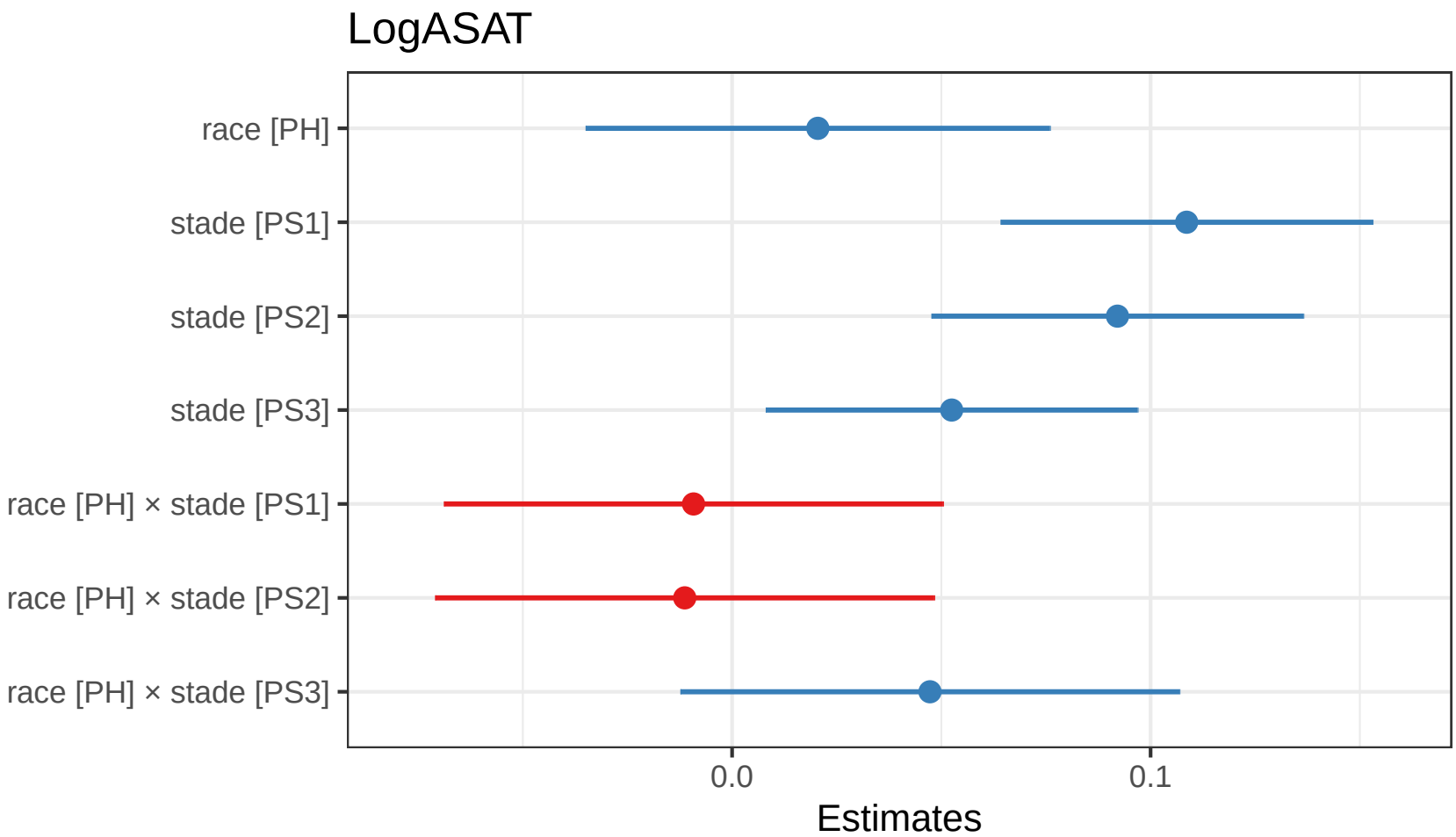
```
confint(mm)
```

```
##                      2.5 % 97.5 %
## .sig01             0.04696 0.0777
## .sigma             0.06755 0.0833
## (Intercept)        1.95992 2.0421
## racePH            -0.03425 0.0750
## stadePS1           0.06499 0.1525
## stadePS2           0.04845 0.1359
## stadePS3           0.00884 0.0963
## racePH:stadePS1 -0.06804 0.0493
## racePH:stadePS2 -0.07014 0.0472
## racePH:stadePS3 -0.01154 0.1058
```

Visualisation des intervalles de confiance sur les effets fixes

ATTENTION, les IC représentés ne sont pas calculés avec la méthode par défaut de `confint()` ("Wald" au lieu de "profile").

```
require(sjPlot)
plot_model(mm) + ylim(-0.08, 0.16)
```



Graphe des résidus

```
plot(mm)
```

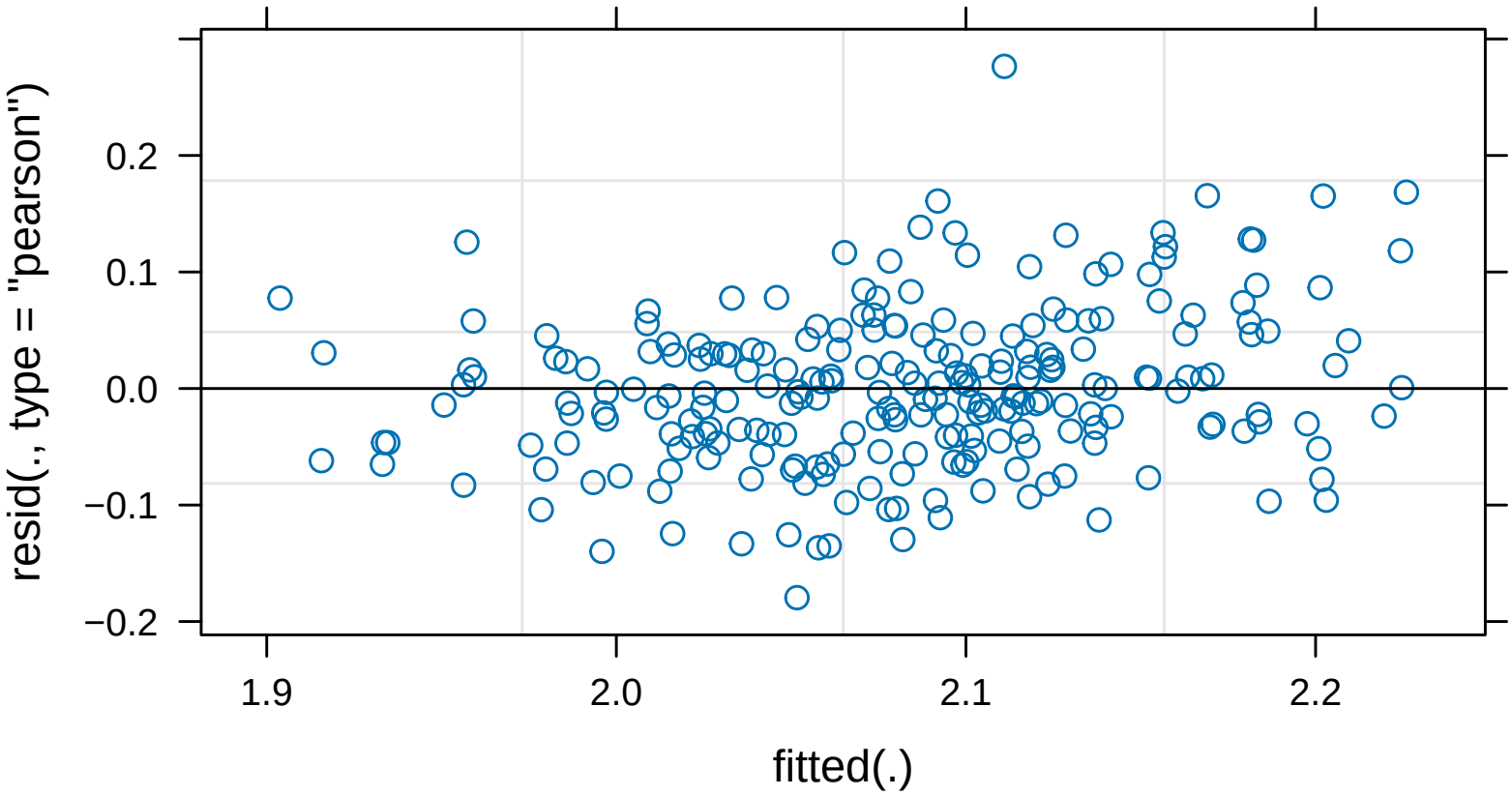


Diagramme quantile-quantile des résidus

```
qqnorm(residuals(mm))
```

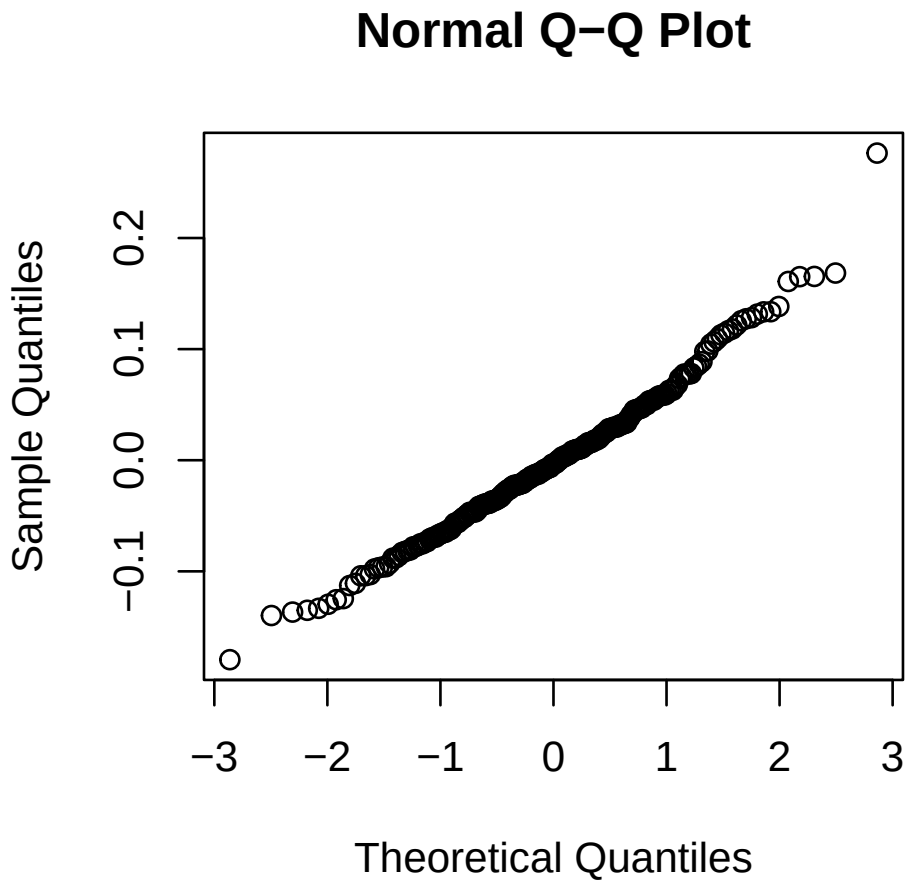
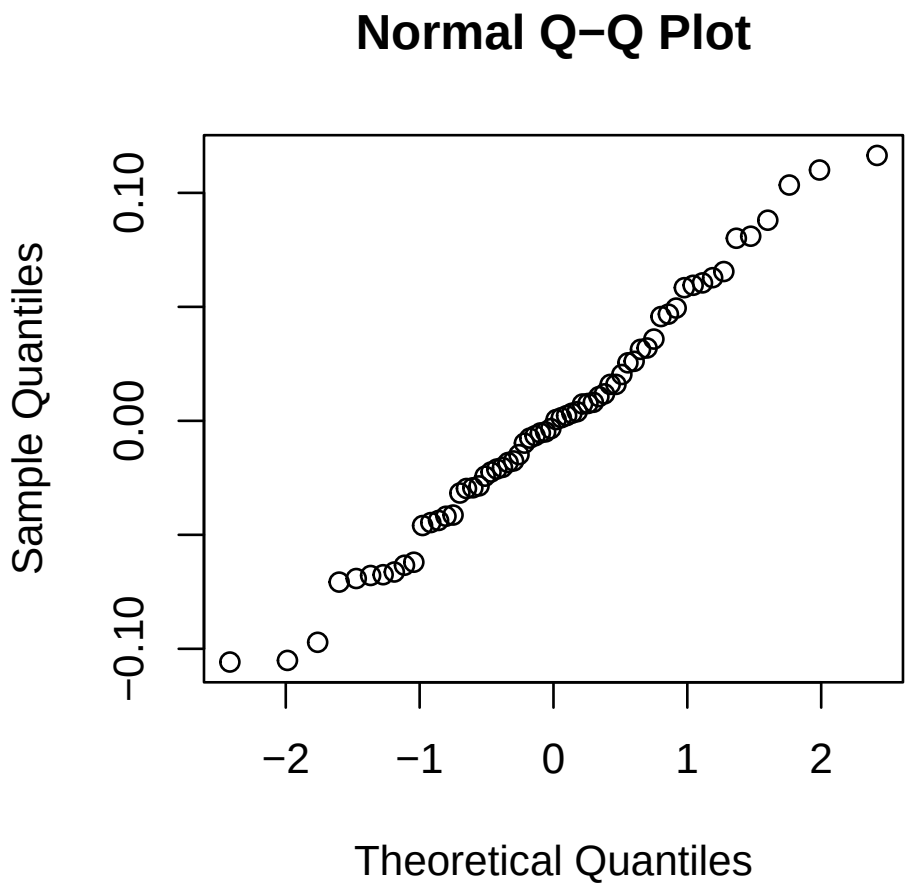


Diagramme quantile-quantile des effets aléatoires liés à la vache

```
qqnorm(ranef(mm)$vache[, 1])
```



Conclusion à partir du modèle

Même si l'on n'exclut pas un effet de la race (ASAT des Prim'holstein légèrement supérieur à l'ASAT des Montbéliardes), et une interaction entre la race et le stade (l'ASAT au stade 3 semble être plus élevé qu'au stade 2 pour les Prim'holstein et plus faible pour les Montbéliardes), le seul effet qui apparaît significatif dans cette étude est celui du stade. Comme il ne semble pas raisonnable d'enlever l'interaction du modèle, les effets fixes sont difficiles à interpréter.

Une possibilité de simplification serait de faire un modèle par race.

Les écarts types intra-individuel et inter-individuel sont respectivement estimés à 0.076 et 0.062.

Exemple de modèle sur la race Prim'holstein

```
mmPH <- lmer(LogASAT ~ stade + (1|vache),
             data = d[d$race == "PH", ])
summary(mmPH)$coefficients
```

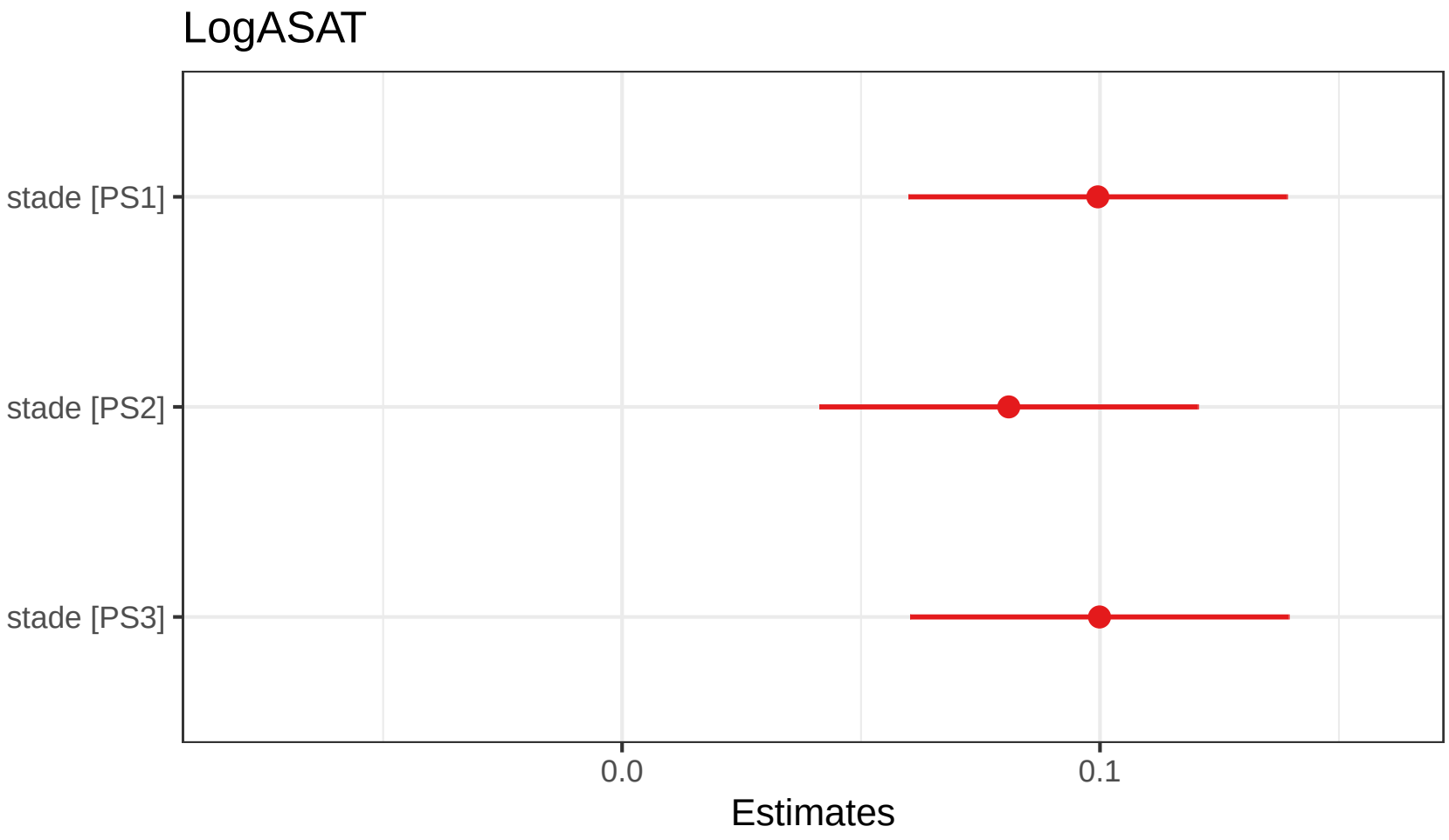
```
##             Estimate Std. Error t value
## (Intercept)   2.0213     0.0189  106.75
## stadePS1      0.0996     0.0200    4.98
## stadePS2      0.0809     0.0200    4.05
## stadePS3      0.0999     0.0200    5.00
```

```
summary(mmPH)$varcor
```

```
##  Groups   Name        Std.Dev.
##  vache    (Intercept) 0.0663
##  Residual             0.0754
```

Visualisation des intervalles de confiance sur les effets fixes pour la race Prim'holstein

```
plot_model(mmPH) + ylim(-0.08, 0.16)
```



Caractérisation d'un exemple à plus de 3 facteurs

Données de surveillance de la fièvre Q chez les caprins, recueillies au cours du temps, par divers labos qui utilisent les trois tests (chaque labo n'utilisant qu'un test), divers lots du test forcément, et avec un effet plaque à prendre en compte, des serums différents (échantillons venant d'animaux pris au hasard, une seule mesure par échantillon).

Si l'on souhaite voir s'il y a une évolution au cours du temps en définissant le facteur année, quels facteurs doit-on prendre en compte pour écrire le modèle ?

Caractérisation d'un exemple à 5 facteurs

- test (fixe à 3 modalités)
- labo (aléatoire imbriqué dans test)
- plaque (aléatoire imbriqué dans labo)
- lot (aléatoire partiellement croisé avec le labo, imbriqué dans test)
- année (fixe, considérée comme qualitative ou quantitative, à voir), avec interaction potentielle entre l'année et le test.

Caractérisation d'un exemple à 5 facteurs - modèle

- test (fixe à 3 modalités)
- labo (aléatoire imbriqué dans test)
- lot (aléatoire partiellement croisé avec le labo, imbriqué dans test)
- année (fixe, considérée comme qualitative ou quantitative, à voir), avec interaction potentielle entre l'année et le test.

Modèle envisageable :

```
(mm <- lmer(ODR ~ test + annee + annee:test +
                (1|labo/plaque) + (1|lot))
```

Conclusion

Conclusion

Vous disposez de toutes les briques de légo pour traiter divers exemples à plus de deux facteurs.
A vous de jouer avec ces briques de légo pour construire des modèles mixtes qui tiennent la route !

- **La référence associée au package lme4** :

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. https://www.jstatsoft.org/article/view/v067i01

- **Une référence pour aller plus loin**

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ, 6, e4794. https://peerj.com/articles/4794/

Méfiez-vous des modèles trop complexes !

Certains ont tendance à écrire des modèles avec toutes les interactions possibles, sans réfléchir à leur sens biologique, ni à la complexité induite du modèle.

- Evitez les écritures du type `Y ~ A*B*C` qui incluent toutes les interactions possibles.
- Construisez un **modèle parcimonieux**, en y mettant juste ce qui est utile pour une bonne description des données en cohérence avec les hypothèses biologiques.
- Faites vos **choix à partir des connaissances biologiques** et des questions que vous vous posez.
- **Regardez toujours les coefficients estimés de votre modèle**. Si vous êtes incapable de les interpréter et/ou que tous sont non significatifs, c'est sans doute que votre modèle est trop complexe.

Aucune méthode automatique ne peut faire cela à votre place !

Les incontournables à mettre dans un article

- **Statistical analysis**
  - use of a **linear mixed-effect model** (with the lme4 R package)
  - **fixed effects** taken into account (and wether each is considered continuous or categorical if it is not obvious)
  - **interactions** included in the model
  - **random effects** taken into account

Mieux vaut éviter de mettre une pseudo formule de code R dans la première partie (pas très explicite) !

- **Results**
  A table with **estimated effects** (with confidence intervals) and/or a forest plot of those estimations.

Mieux vaut éviter les tables d'ANOVA à l'ancienne avec p-value par effet fixe ! (**souvent fausses car difficiles à calculer en cas de plan d'expérience non équilibré** - cas courant - et **pas très informatives**)