Introduction à la sélection

Timothée Flutre (INRA)

25/06/2018

Résumé

Ce document a pour but de présenter brièvement la théorie en génétique quantitative d'intérêt pour la sélection artificielle.

Table des matières

1 Préambule

Ce document a été généré à partir d'un fichier texte au format Rmd utilisé avec le logiciel libre R. Pour exporter un tel fichier vers les formats HTML et PDF, installez le paquet rmarkdown (il va vraisemblablement vous être demandé d'installer d'autres paquets), puis ouvrez R et entrez:

```
library(rmarkdown)
render("intro-sel.Rmd", "all")
```

Il est généralement plus simple d'utiliser le logiciel libre RStudio, mais ce n'est pas obligatoire. Pour plus de détails, lisez cette page.

Le format Rmd permet également d'utiliser le language LaTeX pour écrire des équations. Pour en savoir plus, reportez-vous au livre en ligne.

De plus, ce document nécessite de charger des paquets additionnels (ceux-ci doivent être installés au préalable sur votre machine, via `install.packages("pkg")`):

```
suppressPackageStartupMessages(library(MASS))
```

2 Contexte

Ce document peut surtout être d'intérêt pour les étudiants en génétique quantitative, par exemple ceux suivant l'atelier "Prédiction et Sélection Génomique" organisé et animé par Jacques David et Timothée Flutre depuis 2015.

Les versions du contenu sont gérées avec le logiciel git, et le dépôt central est hébergé sur GitHub.

3 Application

3.1 Modélisation théorique

Concentrons-nous sur un seul caractère d'une espèce donnée, et imaginons une population de I génotypes, chacun ayant une valeur génotypique notée g_i, avec $i \in \{1, \ldots, I\}$. Dans la théorie classique de la génétique quantitative, la valeur génotypique est décomposée en valeurs additive (a_i), de dominance (d_i) et d'épistasie (ζ_i), indépendantes les unes des autres, $g_i = a_i + d_i + \zeta_i$, de telle sorte que a_i corresponde à la part héritable, c'est-à-dire transmise à la descendance, donc particulièrement d'intérêt en sélection, d'où le fait qu'on l'appelle aussi *breeding value*.

L'architecture génétique du caractère est aussi supposée infinitésimale, ce qui permet de la modéliser comme suit: $\forall i,\ g_i \sim \mathcal{N}(0, \sigma_g^2)$. Dans le cas de la valeur additive: $\forall i,\ a_i \sim \mathcal{N}(0, \sigma_a^2)$, où σ_a^2 correspond à la variance génétique additive. De manière similaire pour d_i et ζ_i. Sous forme multivariée, on écrit: $\boldsymbol{a} \sim \mathcal{N}_I(\boldsymbol{0}, \sigma_a^2 A)$, où A est la matrice des relations génétiques additives dont l'espérance, sous les hypothèses de transmission mendélienne, est calculable à partir du pédigrée (A est aussi estimable directement à partir de données de génotypage).

Tant qu'il existe "suffisamment" de variation d'origine génétique additive entre les génotypes ($\sigma_a^2 >> 0$), on peut espérer augmenter la valeur génotypique moyenne de la population au fil des générations en sélectionnant les reproducteurs parmi ceux ayant les meilleurs valeurs additives ($a_i^{(s)}$).

3.2 Modélisation statistique

Or les valeurs génotypiques ne sont pas observables, il faut donc les estimer/prédire à partir d'un échantillon de N observations phénotypiques, notées $\{y_n\}_{1 \leq n \leq N}$, avec $\mathrm{E}[y_n] = \mu_0$ et $\mathrm{Var}[y_n] = \sigma_p^2$. Si l'on suppose aussi que les données sont distribuées selon une Normale, sans covariance entre les erreurs, la vraisemblance s'écrit:

$$\forall n,\ y_n | \mu_0, \sigma_p^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_0, \sigma_p^2)$$

Classiquement, on collecte plusieurs observations (J) pour chaque génotype, ce qui fait que $N = I \times J$ lorsqu'il n'y a pas de données manquantes. En supposant qu'il n'y a pas besoin de prendre compte des corrélations spatio-temporelles entre les observations, la vraisemblance peut alors s'écrire comme:

$$\forall i, j,\ y_{ij} = \mu + g_i + \epsilon_{ij} \text{ avec } \epsilon_{ij} | \sigma^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2) \ \Leftrightarrow\ \forall i, j,\ y_{ij} | \mu, g_i, \sigma^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu + g_i, \sigma^2)$$

Après intégration des $\{g_i\}_{1 \leq i \leq I}$, on obtient:

$$\forall i, j,\ y_{ij} | \mu_0, \sigma_g^2, \sigma^2 \ \sim \mathcal{N}(\mu_0, \sigma_g^2 + \sigma^2)$$

où $\sigma_g^2 + \sigma^2 = \sigma_p^2$.

On est en présence d'un modèle linéaire mixte. Dans le paradigme fréquentiste, l'inférence des composantes de la variance (σ_g^2 et σ^2) peut se réaliser via la méthode *ReML*. Ensuite, le $BLUP$ (resp. $BLUE$) de g_i (resp. μ) peut être calculé à partir des équations de Henderson (MME).

Les composantes de la variance sont également utilisées pour calculer la corrélation entre les valeurs génotypiques et les observations phénotypiques:

$$\rho_{g,y} \ = \ \frac{\mathrm{Cov}[g, y]}{\sigma_g \ \sigma_p} = \frac{\sigma_g^2 + \mathrm{Cov}[g, \epsilon]}{\sigma_g \ \sigma_p}$$

Dans le cas où il n'y a pas de covariance entre génotypes et environnement, on appelle **héritabilité au sens large** (*broad-sense*) le carré de cette corrélation:

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2}$$

L'héritabilité au sens large nous renseigne donc sur la capacité d'un dispositif de collecte de données phénotypiques de nous renseigner sur les valeurs génotypiques des génotypes impliqués.

3.3 D'une génération à l'autre

Considérons la relation entre le phénotype des enfants et le phénotype moyen de leurs deux parents (le "parent moyen"). Pour cela, notons $y_{\text{mère}}$ (resp. $y_{\text{père}}$) une observation phénotypique de la mère (resp. du père) avec y leur moyenne, $y = \frac{y_{\text{mère}} + y_{\text{père}}}{2}$, et y_e une observation phénotypique de leur enfant.

D'après la théorie:

- $y_{\text{mère}} = a_{\text{mère}} + d_{\text{mère}} + \zeta_{\text{mère}} + \epsilon_{\text{mère}}$;
- $y_{\text{père}} = a_{\text{père}} + d_{\text{père}} + \zeta_{\text{père}} + \epsilon_{\text{père}}$;
- $y_e = a_{\text{mère}} + a_{\text{père}} + \epsilon_e$, puisque les valeurs génotypiques additives correspondent justement à ce qui est transmis.

Pour caractériser la relation entre parents et enfants, effectuons la régression linéaire de y_e sur y. La pente de la droite vaut alors: $\beta_{\text{enfants,parents}} = \frac{\mathrm{Cov}[y, y_e]}{\mathrm{Var}[y]}$.

Commençons par la covariance, au numérateur. Celle-ci contient de nombreux termes, mais beaucoup s'annulent sous les hypothèses d'accouplements aléatoires (panmixie), sans sélection ni covariance génotype-environnement ni transmission d'effets environnementaux. Au final:

$$\mathrm{Cov}[y, y_e] = \mathrm{Cov}\left[\left(\frac{a_{\text{mère}} + a_{\text{père}}}{2}\right), (a_{\text{mère}} + a_{\text{père}})\right] = \frac{\mathrm{Var}[a_{\text{mère}}] + \mathrm{Var}[a_{\text{père}}]}{2}$$

En panmixie, la variance génétique totale dans la population est la somme des variances maternelle et paternelle: $\sigma_a^2 = \mathrm{Var}[a_{\text{mère}}] + \mathrm{Var}[a_{\text{père}}]$.

Passons maintenant à la variance du parent moyen, au dénominateur. En panmixie, la covariance entre phénotypes maternel et paternel est nulle, d'où: $\mathrm{Var}[y] = \frac{\mathrm{Var}[y_{\text{mère}}] + \mathrm{Var}[y_{\text{père}}]}{4}$. De plus, en supposant que la variance phénotypique ne dépende pas du sexe: $\mathrm{Var}[y] = \frac{\sigma_p^2}{2}$.

Au final, la pente de la droite vaut:

$$\beta_{\text{enfants,parents}} = \frac{\sigma_a^2}{\sigma_p^2} = h^2$$

où h^2 est appelée **héritabilité au sens strict** (*narrow-sense*).

Ceci montre que les mesures phénotypiques sur des génotypes apparentés nous renseignent sur la variance génétique additive d'une population.

Notons que h^2 correspond aussi à la corrélation entre observation phénotypique et valeur génotypique additive au sein d'un individu: $\rho_{a,y} = \frac{\text{Cov}[a,y]}{\sigma_a \sigma_p} = \frac{\sigma_a^2}{\sigma_a \sigma_p} = h$.

4 Exemple

Voici un exemple concret qui servira dans la suite, pour lequel on se limite au cas d'une architecture purement additive ($g_i = a_i$). Comme ce document traite de sélection et non d'inférence, on se limite aussi au cas où $J = 1$.

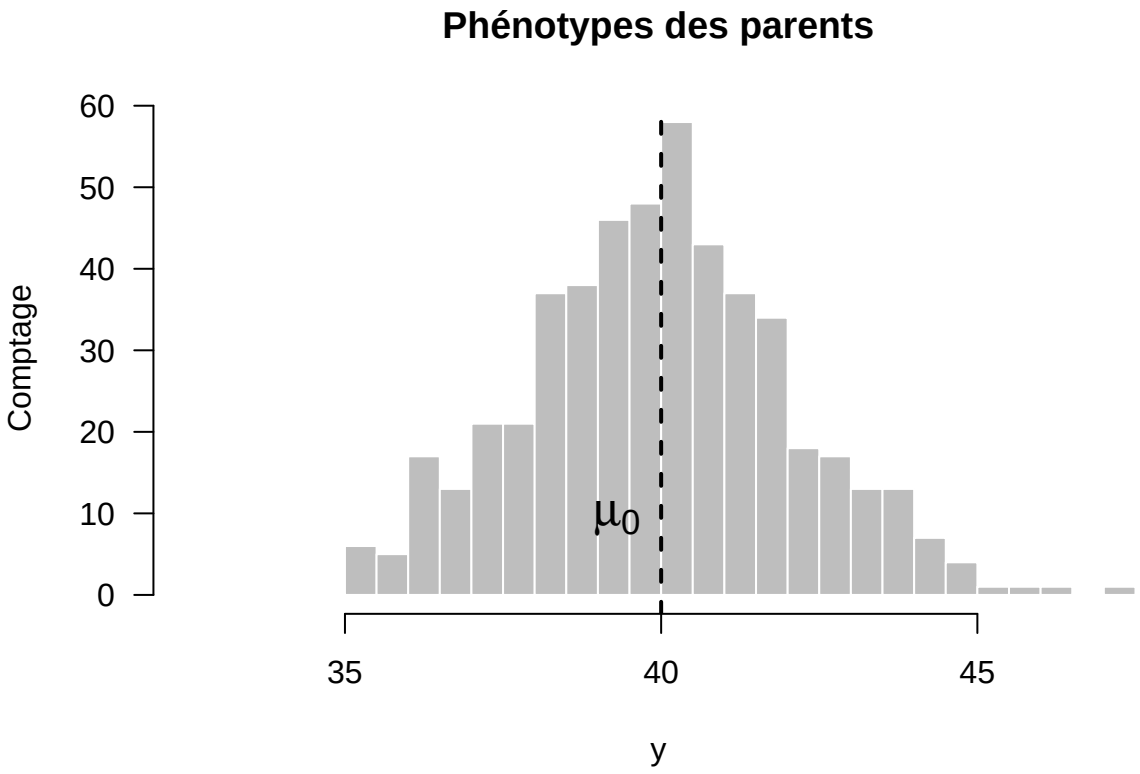
De plus, supposons que la relation entre génotypes des parents et des enfants est linéaire. S'il n'y a pas de sélection, c'est-à-dire que tout génotype peut être parent, les moyenne et variance phénotypiques à la génération des enfants sont égales à celles des parents, μ_0 et σ_p^2. Comme la pente de la droite, $\beta_{\text{enfants,parents}}$, correspond à h^2, on peut en déduire la covariance entre observations phénotypiques des parents et enfants, et ainsi simuler toutes les observations phénotypiques via une loi Normale bivariée.

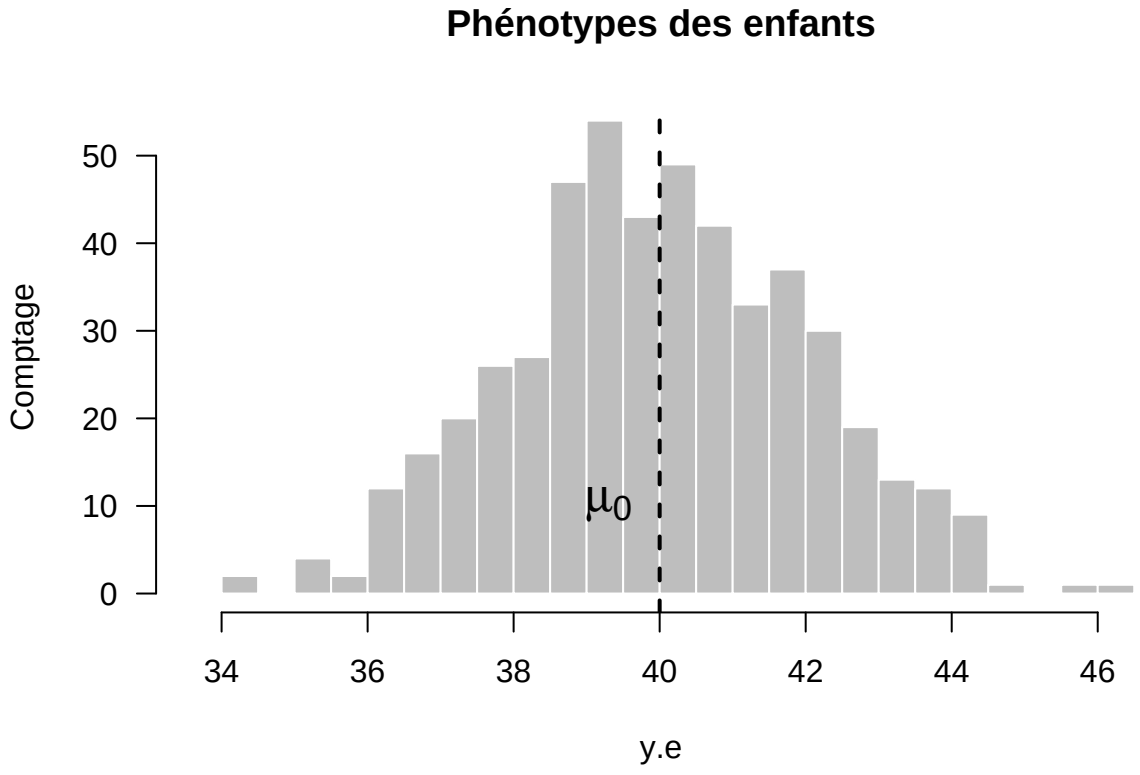
```
set.seed(1859)
I <- 500
J <- 1
N <- I * J
mu.0 <- 40
mean.midparents <- mu.0
mean.offsprings <- mu.0
sigma.a2 <- 3
sigma2 <- 1
var.midparents <- sigma.a2 + sigma2
var.offsprings <- sigma.a2 + sigma2
(h2 <- sigma.a2 / (sigma.a2 + sigma2))
```

```
## [1] 0.75
```

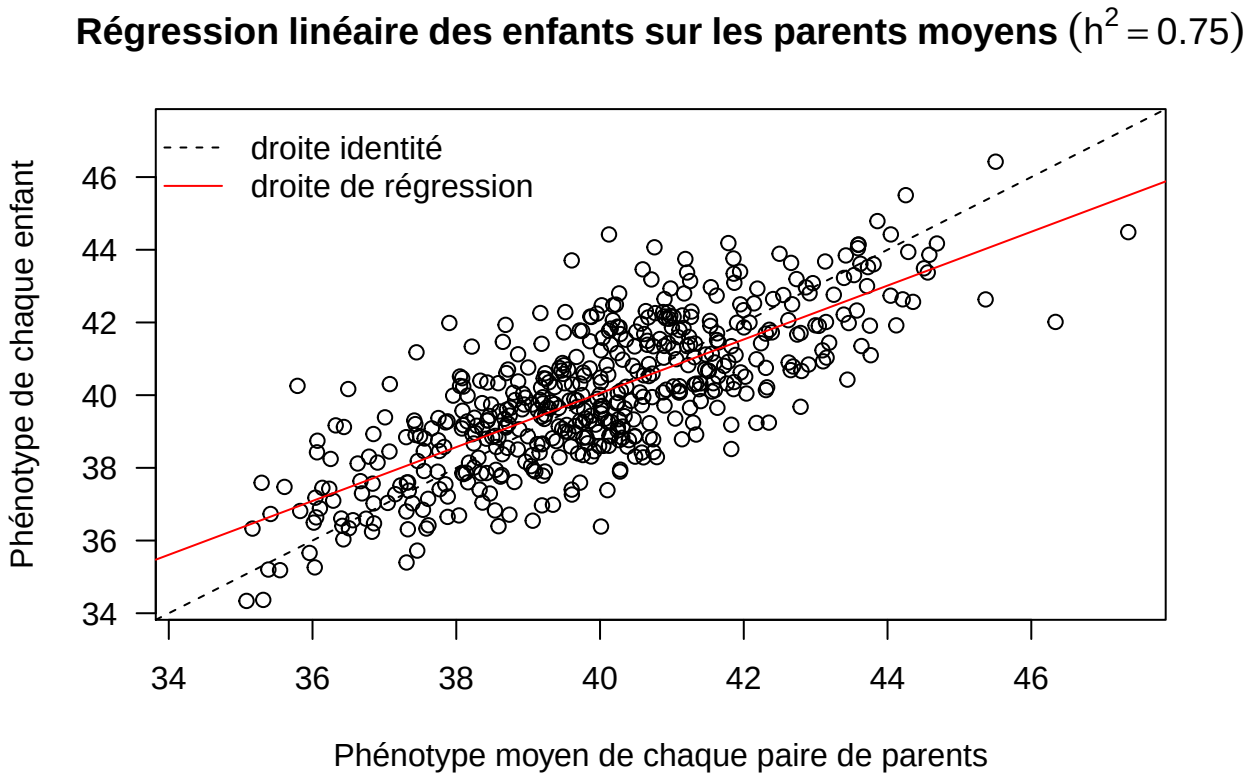
```
covar.midpar.off <- h2 * var.midparents
Sigma <- matrix(c(var.midparents, covar.midpar.off,
                  covar.midpar.off, var.offsprings),
                nrow=2, ncol=2)
all.y <- mvrnorm(n=N, mu=c(mean.midparents, mean.offsprings),
                 Sigma=Sigma)
y <- all.y[,1]   # mid-parents
y.e <- all.y[,2] # offsprings
```

Voici les histogrammes des vraies données:





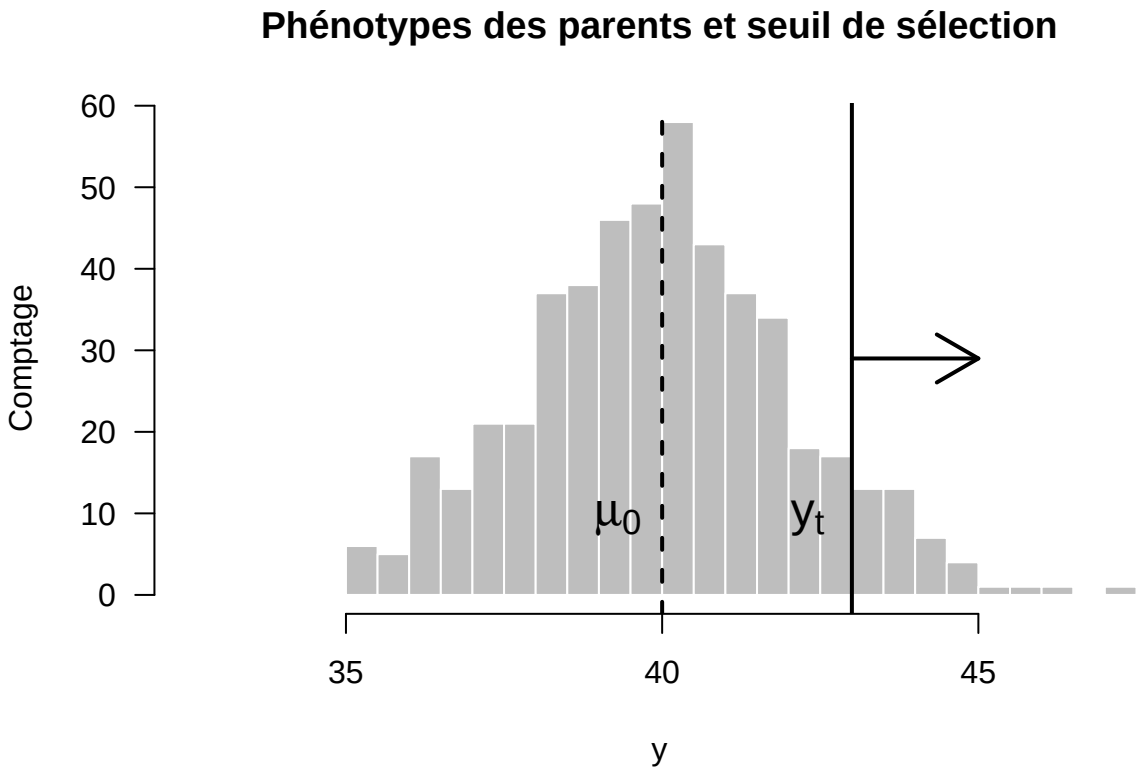
Voici la régression des enfants sur les parents:



5 Sélection par troncation

La sélection s'applique communément par **troncation**, c'est-à-dire en choisissant un seuil phénotypique y_t au-dessus duquel les génotypes sont sélectionnés, c'est-à-dire autorisés/choisis pour se reproduire (par croisement).

Dans l'exemple, fixons le seuil y_t à 43:



6 Différentiel de sélection

A la génération initiale, la moyenne phénotypique est notée μ_0 avant sélection, et $\mu^{(s)}$ après (mais avant reproduction). L'une des mesures possibles de la pression de sélection appliquée est le **différentiel de sélection**:

$$S = \mu^{(s)} - \mu_0$$

Dans l'exemple:

```
sum(is.sel <- (y >= y.t))
```

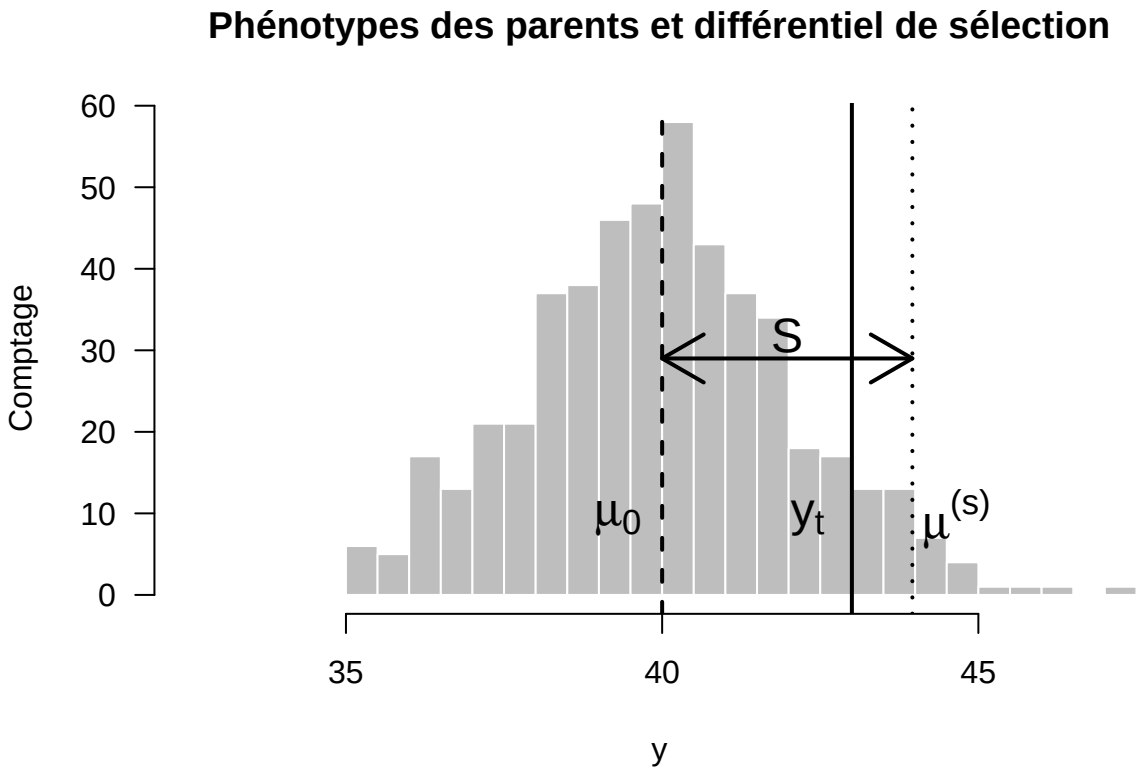
```
## [1] 41
```

```
(mu.s <- mean(y[is.sel]))
```

```
## [1] 44
```

```
(S <- mu.s - mu.0)
```

```
## [1] 3.96
```



Le seuil y_t est relié à la fraction des génotypes sélectionnés, α. Dans l'exemple:

```
(alpha <- sum(is.sel) / length(y))
```

```
## [1] 0.082
```

7 Intensité et taux de sélection

Le désavantage du différentiel de sélection S est de dépendre de l'unité de mesure du phénotype. Pour comparer la sélection sur différents caractères, il est donc recommandé de travailler avec une valeur standardisée, l'**intensité de sélection**, notée i (à ne pas confondre avec l'indice i du modèle statistique):

$$i = \frac{S}{\sigma_p}$$

Dans le cas où les valeurs utilisées pour la sélection (ici les valeurs génotypiques $\{g_i\}_{1 \leq i \leq I}$) suivent une loi Normale, un autre avantage de l'intensité i est de permettre d'approximer sa relation avec le **taux de sélection** α:

$$i = \frac{z}{\alpha}$$

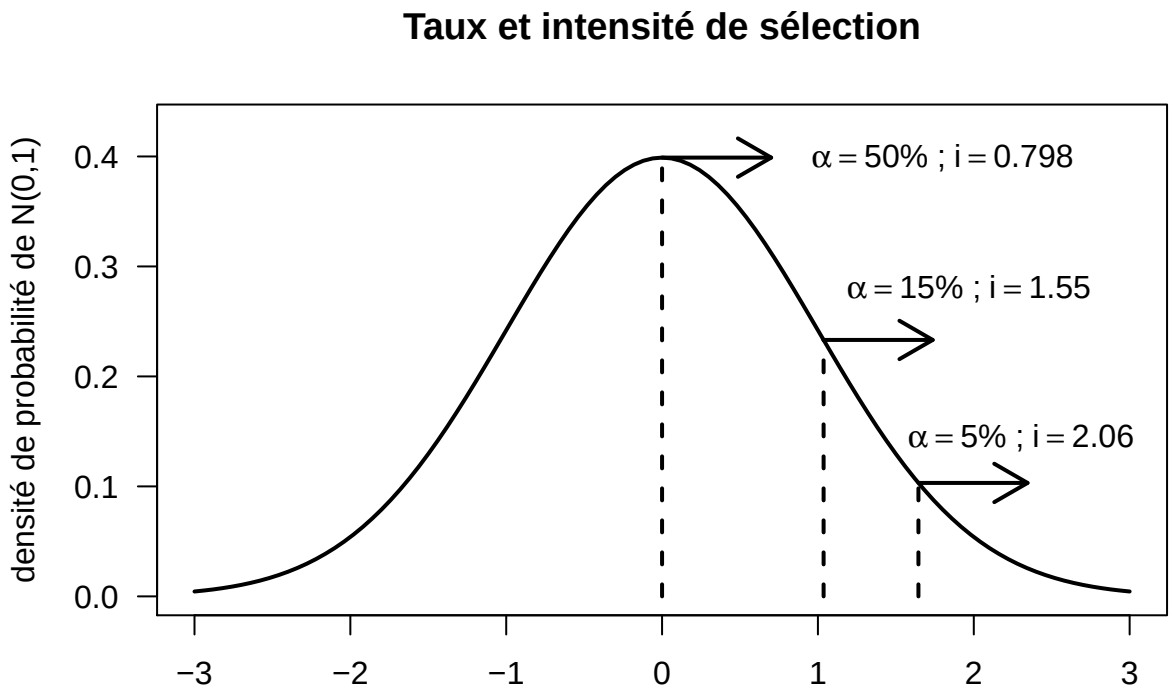
où z est l'ordonnée du point de la loi Normale centrée réduite ayant pour abscisse le seuil correspondant à α.

Il existe des tables contenant les valeurs de i pour différentes valeurs de α, mais on peut facilement les recalculer:

```
alpha <- c(0.01, 0.05, 0.1, 0.15, 0.2, 0.5) # from 1% to 50%
z <- qnorm(p=alpha, mean=0, sd=1, lower.tail=FALSE)
```

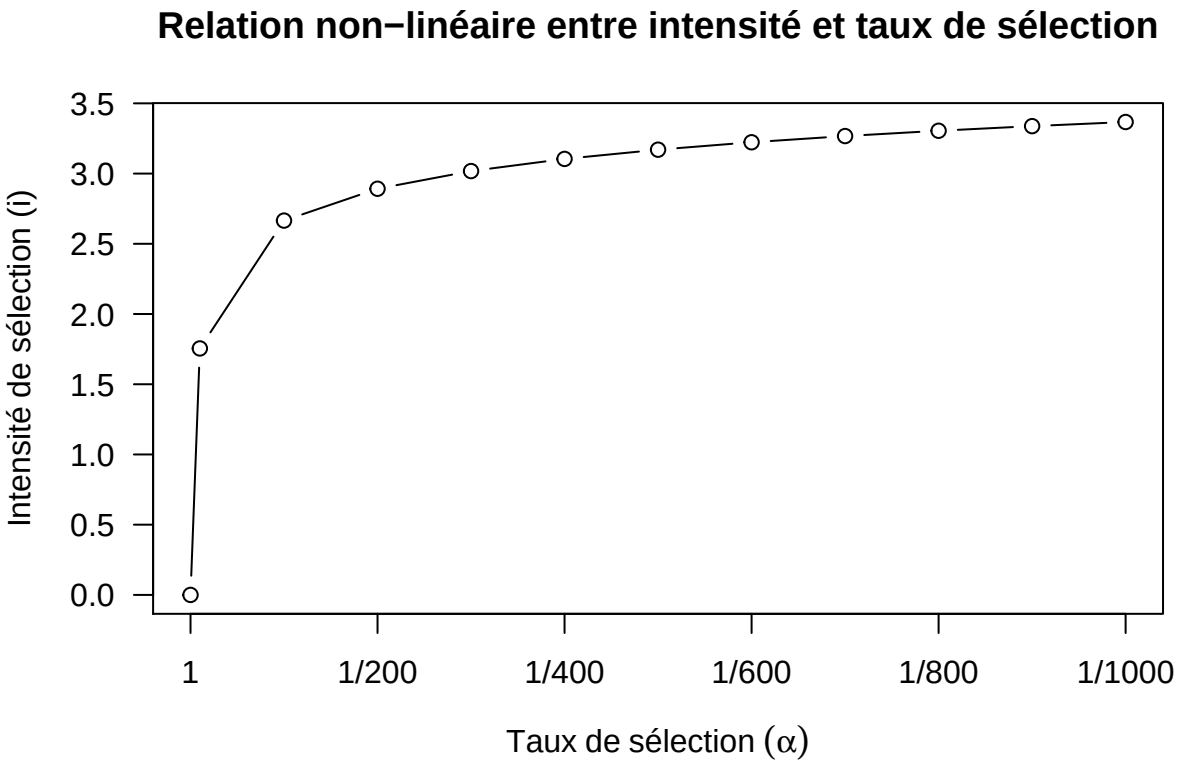
```r
phi.z <- dnorm(x=z, mean=0, sd=1)
(i <- phi.z / alpha)
```

```
## [1] 2.665 2.063 1.755 1.554 1.400 0.798
```



Attention cependant au fait que la précision de cette approximation diminue avec la taille de la population, I, notamment quand $I < 100$.

Remarquez aussi que l'intensité de sélection n'augmente pas linéairement en fonction du taux de sélection:



8 Réponse à la sélection

L'héritabilité au sens strict, h^2, étant inférieure (ou égale) à 1, la droite de régression des enfants sur les parents n'est donc généralement pas aussi pentue que la droite identité. C'est cette observation qui a amené Galton a parlé de "régression".

Or, comparée à la moyenne des parents sélectionnés, $\mu^{(s)}$, la moyenne de leurs enfants, notée μ_1, est plus élevée, ce qui amène à définir la **réponse à la sélection** (aussi appelé **gain génétique**):

$$R = \mu_1 \ - \ \mu_0$$

Dans l'exemple:

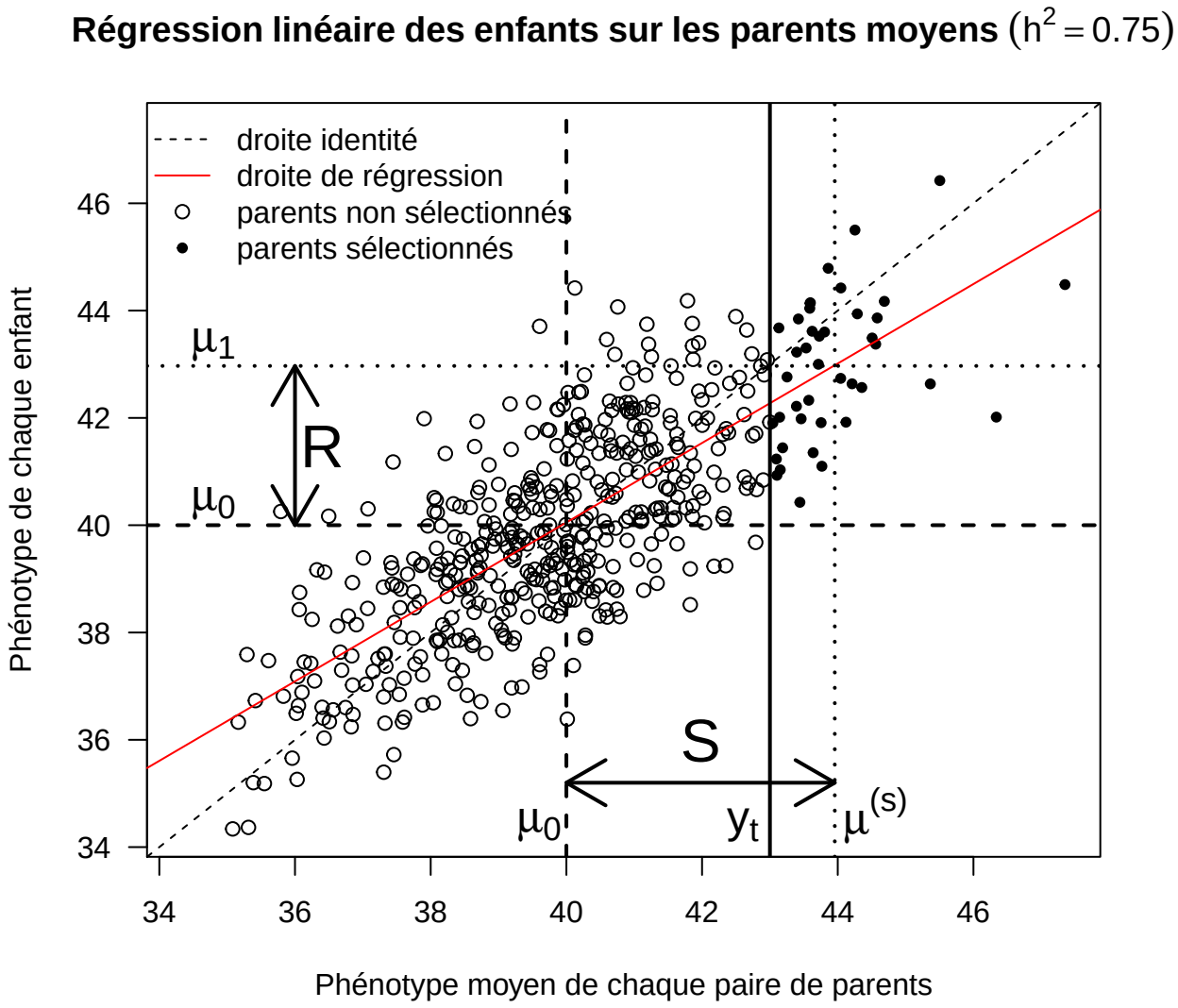
```
(mu.1 <- mean(y.e[is.sel]))
```

```
## [1] 43
```

```
(R <- mu.1 - mu.0)
```

```
## [1] 2.97
```

Si l'on résumé tout sur un seul et même graphique, cela donne:



On voit que le changement, dû à la sélection, de moyenne phénotypique au sein des parents (S) induit bien un changement de moyenne phénotypique au sein de leurs enfants (R):

$$\beta_{\text{enfants,parents}} = \frac{\mu_1 - \mu_0}{\mu^{(s)} - \mu_0} \Leftrightarrow h^2 = \frac{R}{S}$$

Cette équation combine une information d'hérédité (h^2) avec une information de sélection (S) pour prédire le changement d'une génération à l'autre (R). Cette relation est généralement appelée l'**équation du sélectionneur**:

$$R = h^2 \, S$$

De plus, comme $S = i\,\sigma_p$, on peut écrire: $R = i\,h^2\,\sigma_p = i\,h\,\sigma_a$, où $h = \frac{\sigma_a}{\sigma_p}$. On peut aussi voir h comme la corrélation entre les observations phénotypiques $\{y_n\}$ et les valeurs génotypiques additives, $\{a_i\}$, d'où le fait que h soit aussi notée r pour indiquer qu'elle mesure la précision (*accuracy*) avec laquelle les observations phénotypiques permettent d'approximer les valeurs génotypiques additives:

$$R = i\;r\;\sigma_a$$

En bref: **réponse = intensité × précision × variance génétique additive**.

9 Optimisation de la sélection

La réponse à la sélection (par unité de temps) dépend de l'intensité de sélection, de la précision de prédiction des valeurs génotypiques additives et de la variance génétique additive. L'intensité de sélection dépend à son tour du nombre de génotypes candidats testés et de la fraction de sélectionnés. La précision de prédiction dépend quant à elle du nombre de génotypes testés, mais aussi du nombre de sites, d'années et de réplicats.

Si l'on veut un nombre minimum de génotypes remplissant le critère de sélection, un nombre suffisant de génotypes testés est nécessaire. De plus, un nombre suffisant d'observations est requis pour évaluer précisement les valeurs génotypiques. A budget constant, il est donc nécessaire d'optimiser les paramètres d'un programme de sélection, et donc de mettre au point une stratégie.

Si on sélectionne très fortement dès la première génération en ne permetant qu'à un tout petit nombre de génotypes de se reproduire, on va certes augmenter fortement la moyenne de la génération suivante, mais on va aussi diminuer drastiquement la variance génétique au sein de celle-ci. Un équilibre doit donc être trouvé entre augmenter la moyenne de génération en génération, sans pour autant perdre trop de variance.

10 Références

D'accès gratuit et en français:

— Minvielle (1990): "Principes d'amélioration génétique des animaux domestiques"

Payant et en anglais, mais référence incontournable:

— Lynch et Walsh (1998): "Genetics and analysis of quantitative traits"

11 Annexe

```
print(sessionInfo(), locale=FALSE)
```

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.4 LTS
##
## Matrix products: default
## BLAS: /usr/lib/openblas-base/libblas.so.3
## LAPACK: /usr/lib/libopenblasp-r0.2.18.so
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] MASS_7.3-50   knitr_1.20    rmarkdown_1.9
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.4.4  backports_1.1.2 magrittr_1.5    rprojroot_1.3-2
##  [5] tools_3.4.4     htmltools_0.3.6 yaml_2.1.19     Rcpp_0.12.17
##  [9] stringi_1.2.2   methods_3.4.4   stringr_1.3.1   digest_0.6.15
## [13] evaluate_0.10.1
```