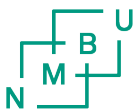
第一天：育种值估计概览

于希江

挪威生命科学大学畜牧与水产系

二〇一八 • 十一月
青岛



目录

育种和育种值

从加性遗传模型理解育种值
- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计
- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

动物生产

种 育种和遗传。

料 饲料配方和工艺。

病 卫生防疫，疫病控制。

管 动物生产系统管理。

动物生产

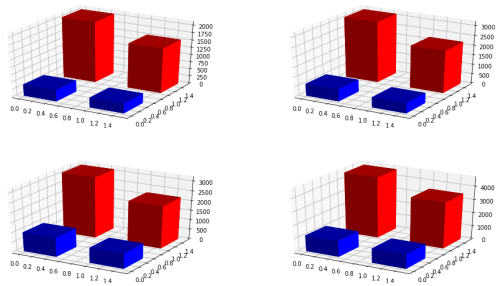
种 育种和遗传。

料 饲料配方和工艺。

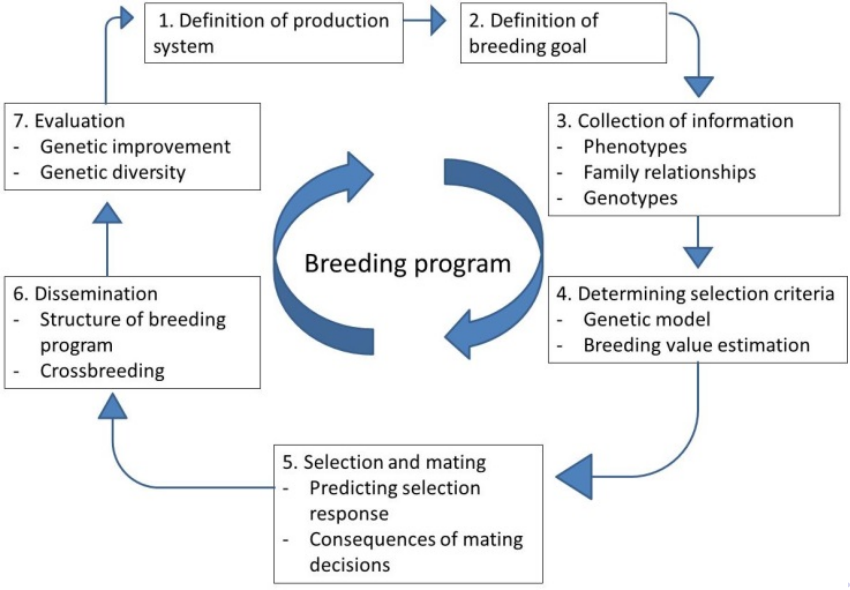
病 卫生防疫，疫病控制。

管 动物生产系统管理。

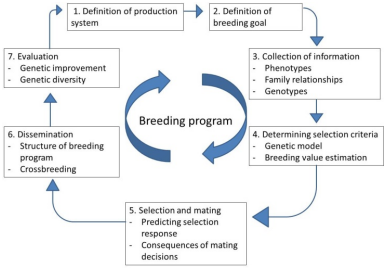
- 肉鸡 1957，2001 品系交叉饲喂同期饲料胴体重对比（公鸡，克；43，57，71，和 85 天结果。Havenstein et al，2003.）：



育种的流程


1. Definition of production system
2. Definition of breeding goal
3. Collection of information
- Phenotypes
- Family relationships
- Genotypes
4. Determining selection criteria
- Genetic model
- Breeding value estimation
5. Selection and mating
- Predicting selection response
- Consequences of mating decisions
6. Dissemination
- Structure of breeding program
- Crossbreeding
7. Evaluation
- Genetic improvement
- Genetic diversity
Breeding program

育种的流程



- 摘自 https://wiki.groenkennisnet.nl/display/TAB

- 确定生产系统
- 确定育种目标
- 收集信息
  - 表型
  - 系谱资料
  - 基因型
- 确定选择标准
  - 遗传模型
  - 育种值估计

- 选择和交配
  - 预测选择反应
  - 交配结果的预后
- 传播
  - 育种项目的结构
  - 杂交
- 评估
  - 遗传改良
  - 遗传多样性

育种和育种值

育种的定义

指在一个群体中，在控制近交的情况下获得遗传进展。

有关遗传进展

$$\Delta G = \frac{ir\sigma_a}{L}$$

育种和育种值

育种的定义

指在一个群体中，在控制近交的情况下获得遗传进展。

有关遗传进展

$$\Delta G = \frac{ir\sigma_a}{L}$$

育种值定义（之一）

一个特定群体中，如果一个个体可以与该群体中的个体随机交配产生后代，那么这些后代在某个性状上的表型的平均值与群体平均数差异的二倍，就是该个体在该群体中关于该性状的育种值。

育种和育种值

育种的定义

指在一个群体中，在控制近交的情况下获得遗传进展。

有关遗传进展

$$\Delta G = \frac{ir\sigma_a}{L}$$

育种值定义（之一）

一个特定群体中，如果一个个体可以与该群体中的个体随机交配产生后代，那么这些后代在某个性状上的表型的平均值与群体平均数差异的二倍，就是该个体在该群体中关于该性状的育种值。

育种值的估计

$$\begin{aligned} P &= G + E \\ &= \overbrace{A + D + I}^{G} + \overbrace{E_C + E_S}^{E} \\ \overline{P} &= \overline{G} \quad \Leftarrow \overline{E} = 0 \end{aligned}$$

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

加性遗传模型

$$
\begin{aligned}
P &= G + E \\
&= \overbrace{A + D + I}^{G} + \overbrace{E_C + E_S}^{E} \\
e_i &\overset{\text{iid}}{\sim} N(0, \sigma^2)
\end{aligned}
$$

加性遗传模型

$$
\begin{aligned}
P &= G + E \\
&= \overbrace{A + D + I}^{G} + \overbrace{E_C + E_S}^{E} \\
e_i &\overset{\text{iid}}{\sim} N(0, \sigma^2)
\end{aligned}
$$

- 有了该模型，我们可以：利用系谱或者基因组数据估计育种值
  - 模拟产生数据集 D
    - 包括系谱、基因型和表现型
  - 尝试用若干种统计方法分析 D
    - 估计育种值
    - 利用常用的统计软件如（Jupyter + Python）
    - 多次模拟，评估估计效果
    - 选项：估计遗传力

理解统计模型和一般估计方法

- 要了解的概念
  - 随机数
  - 模型/建模
  - 方程组
  - 普通最小二乘

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

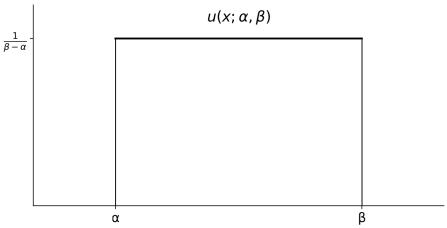
线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

随机数

- 真随机数，如：
  - 一些物理现象
    - 盖革计数器相邻两个声音的间隔时间
    - 大气噪声（例如：**https://random.org**）
  - 现在大部分 Linux/Unix 机器：/dev/urandom
- 伪随机数，如：
  - 线性同余发生器
    - $N_{i+1} = (A \cdot N_i + B) \mod M$
    - 如 glibc：$N_{i+1} = (1103515245 \times N_i + 12345) \mod 2^{32}$
  - mt19937

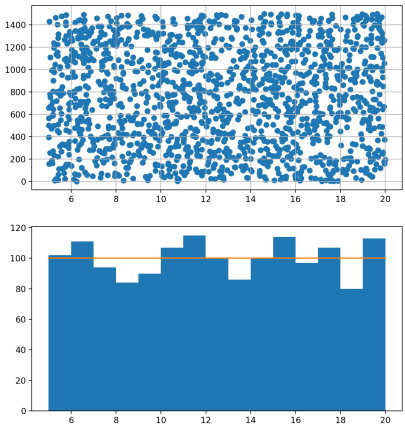
均匀分布



- 均匀分布概率密度函数

$$u(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \text{for } \alpha < x < \beta \\ 0 & \text{elsewhere} \end{cases}$$

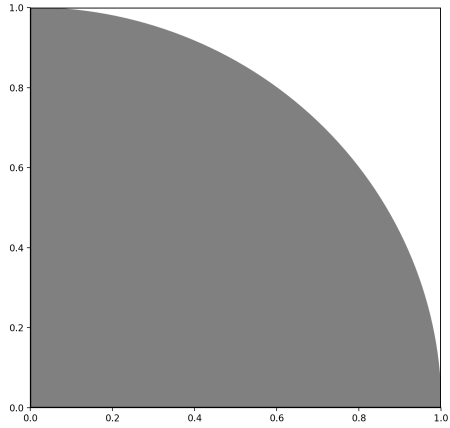
均匀分布随机数示例



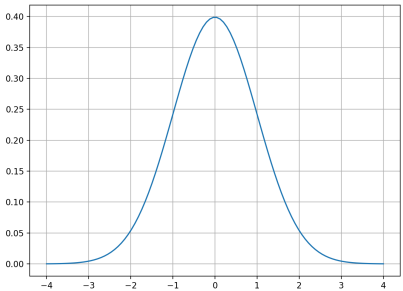
- $1,500$ 均匀分布随机数 $r \sim U(5, 20)$
  - 上图：$1,500$ 点的散点图
  - 下图：柱形图

蒙特·卡罗模拟一例

- 模拟用的均匀分布随机数需要
  - 发生速度快
  - 长周期（mt19937：$2^{19937}-1$）
  - 相邻的 n 个数字应尽可能均匀分布于每维 α 到 β 的 n 维空间。
  - 其它
- 如若 $x_i \sim U(0,1)$ 则点 (x_{2i}, x_{2i-1}) 均匀分布在以 $(0,\ 0)$ 到 $(1,\ 1)$ 为对角线的正方形中。
- 于是，落入右图阴影部分的面积是 $\pi/4$。
- 见 day-1-some-random-numbers.ipynb



正态分布

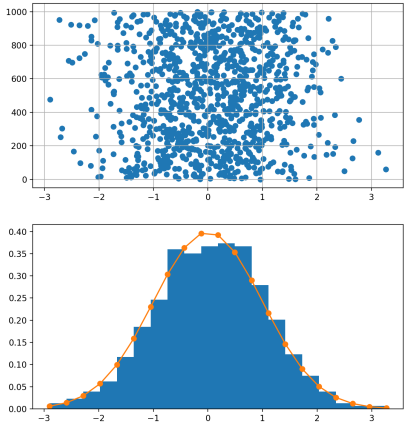


- 正态分布概率密度函数

$$n(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\quad x\in(-\infty,\infty)$$

正态分布随机数示例

- 1,000 个标准正态分布随机数 $r \sim U(5, 20)$
  - 上图：散点图
  - 下图：柱形图



目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

统计模型示例

单因素方差分析模型

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

统计模型示例

单因素方差分析模型

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

其中：

- y_{ij} 处理第 i 水平，第 j 重复的观测值
- μ 整体平均
- α_i 第 i 水平效应
- e_{ij} 每个观测值的随机残差，一般假定 $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$

统计模型示例

“基因组选择”模型

$$y_i = \mu + \sum_{j=1}^{N_{\text{marker}}} x_{ij} b_j + e_i$$

统计模型示例

“基因组选择”模型

$$y_i = \mu + \sum_{j=1}^{N_{\text{marker}}} x_{ij} b_j + e_i$$

其中：

- y_i 个体 i 的表型观测值
- μ 群体平均
- x_{ij} 个体 i 第 j 座位的基因型
- b_i 第 j 座位的基因型值
- e_i 个体 i 的随机残差效应，一般假定独立、同质服从正态分布。

统计模型示例

“基因组选择”模型

$$y_i = \mu + \sum_{j=1}^{N_{\text{marker}}} x_{ij} b_j + e_i$$

其中：

y_i 个体 i 的表型观测值

μ 群体平均

x_{ij} 个体 i 第 j 座位的基因型

b_i 第 j 座位的基因型值

e_i 个体 i 的随机残差效应，一般假定独立、同质服从正态分布。

$\sum_{j=1}^{N_{\text{marker}}} x_{ij} b_j$ 育种值

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

方程组

今有雉兔同笼，上有三十五头，
下有九十四足。问雉兔各几何。

——《孙子算经》

- 设有 x_1 雉，x_2 兔，则有二元一次方程组：

$$\begin{aligned} x_1 &+ x_2 &= 35 \\ 2x_1 &+ 4x_2 &= 94 \end{aligned}$$

方程组

今有雉兔同笼，上有三十五头，
下有九十四足。问雉兔各几何。

——《孙子算经》

- 设有 x_1 雉，x_2 兔，则有二元一次方程组：

$$\begin{aligned} x_1 &+ x_2 &&= 35 \\ 2x_1 &+ 4x_2 &&= 94 \end{aligned}$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 35 \\ 94 \end{bmatrix}$$

方程组

今有雉兔同笼，上有三十五头，
下有九十四足。问雉兔各几何。

——《孙子算经》

- 设有 x_1 雉，x_2 兔，则有二元一次方程组：

$$\begin{aligned} x_1 &+ x_2 &= 35 \\ 2x_1 &+ 4x_2 &= 94 \end{aligned}$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 35 \\ 94 \end{bmatrix}$$

- 若干概念：
  - 矩阵、向量、维数/度、方阵、转置、对称阵、秩、满秩、逆矩阵
  - 以上：$\mathrm{A} \cdot \mathrm{x} = \mathrm{b} \Rightarrow \mathrm{x} = \mathrm{A}^{-1} \cdot \mathrm{b} = \begin{bmatrix} 2 & -0.5 \\ -1 & 0.5 \end{bmatrix} \begin{bmatrix} 35 \\ 94 \end{bmatrix} = \begin{bmatrix} 23 \\ 12 \end{bmatrix}$

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

普通最小二乘 OLS

- 简单如单因子方差分析：
  - $y_{ij} = \mu + \alpha_i + e_{ij}$
    - 一个观测值一个方程
    - e_{ij} 在每个方程中都有，且各不相同
    - 整个方程组还包括总的水平数和 1 个群体平均
    - 未知数个数超过方程个数
  - $\Rightarrow$ 方程无解

普通最小二乘 OLS

- 简单如单因子方差分析：
  - $y_{ij} = \mu + \alpha_i + e_{ij}$
    - 一个观测值一个方程
    - e_{ij} 在每个方程中都有，且各不相同
    - 整个方程组还包括总的水平数和 1 个群体平均
    - 未知数个数超过方程个数
  - $\Rightarrow$ 方程无解

- 最小二乘估计
  - 找到这样一组参数 θ 的估计值 $\hat{\theta}$，使得：
    - 残差 e，即估计值 $\hat{y}$ 与观测值 y 之差，的平方和最小。
    - 用式子表示：

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (f_\theta(x_i) - y_i)^2$$

  - 亦即，观测值与估计值之间的欧几里得距离最小。

普通最小二乘 OLS

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (f_\theta(x_i) - y_i)^2$$

- $\hat{\theta} = \arg\min_{\theta} L$: Arguments of the minimum of function L
- 二次型有解析最小值
- 容易推导和计算
- 鲁棒、稳健
- 属于一种优化估计

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

优化估计

什么是优化估计

选择模型的参数 $\hat{\theta}$，使得模型对观测值有最优解释。

$$\hat{\theta} = \arg\min_{\theta} L_i(f_\theta(x_i, \theta), y_i)$$

优化估计

什么是优化估计

选择模型的参数 $\hat{\theta}$，使得模型对观测值有最优解释。

$$\hat{\theta} = \arg\min_{\theta} L_i(f_\theta(x_i, \theta), y_i)$$

三个概念

模型、目标、优化

最小二乘的优化估计

$$\hat{\theta} = \arg\min_{\theta} \sum_{i-1}^{n} (f_\theta(x_i) - y_i)^2$$

- 模型

$$y_i = \sum_{j=1}^{m} x_{ij}\theta_j + e_i$$

- 模型

$$y_i = \sum_{j=1}^{m} x_{ij}\theta_j + e_i$$

- 方程组表示

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

- 或者

$$\mathrm{y} = \mathrm{X}\theta + \mathrm{e}$$

优化估计的目标和优化

目标

令 $\sum(e_i^2)$ 最小。

或者说，我们要寻找这样的参数估计值 $\hat{\theta}$，它们使得 $\sum(y_i - \mathrm{x}\theta)^2$，或者 $(\mathrm{y} - \mathrm{X}\theta)'(\mathrm{y} - \mathrm{X}\theta)$，最小。

优化估计的目标和优化

目标

令 $\sum(e_i^2)$ 最小。

或者说，我们要寻找这样的参数估计值 $\hat{\theta}$，它们使得 $\sum(y_i - \mathrm{x}\theta)^2$，或者 $(\mathrm{y} - \mathrm{X}\theta)'(\mathrm{y} - \mathrm{X}\theta)$，最小。

优化

将左边二次式对 θ 的一阶导数为零时有唯一解。

示例 — 当 θ 中仅含一个参数 b

优化函数 L

$$
\begin{aligned}
L &= \sum_i^n (y_i - x_i b)^2 \\
&= \sum_i^n (y_i^2 - 2x_i y_i b + x_i^2 b^2)
\end{aligned}
$$

示例 — 当 θ 中仅含一个参数 b

优化函数 L

$$\begin{aligned} L &= \sum_i^n (y_i - x_i b)^2 \\ &= \sum_i^n (y_i^2 - 2x_i y_i b + x_i^2 b^2) \end{aligned}$$

对 b 求导

$$L' = \sum_i^n (2x_i^2 b - 2x_i y_i)$$

令优化函数一阶导数 $L' = 0$

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

示例 — 当 θ 中仅含两个参数b_1, b_2

优化函数 L

$$
\begin{aligned}
L &= \sum_i^n (y_i - x_{i1}b_1 - x_{i2}b_2)^2 \\
&= \sum_i^n (y_i^2 + x_{i1}^2 b_1^2 + x_{i2}^2 b_2^2 \\
&\quad + 2x_{i1}x_{i2}b_1 b_2 - 2x_{i1}y_i b_1 - 2x_{i2}y_i b_2)^2
\end{aligned}
$$

示例 — 当 θ 中仅含两个参数b_1, b_2

优化函数 L

$$\begin{aligned} L &= \sum_i^n (y_i - x_{i1}b_1 - x_{i2}b_2)^2 \\ &= \sum_i^n (y_i^2 + x_{i1}^2 b_1^2 + x_{i2}^2 b_2^2 \\ &\quad + 2x_{i1}x_{i2}b_1b_2 - 2x_{i1}y_i b_1 - 2x_{i2}y_i b_2)^2 \end{aligned}$$

分别对 b_1 和 b_2 求导

$$\begin{cases} L'_{b_1} &= \sum (2x_{i1}^2 b_1 + 2x_{i1}x_{i2}b_2 - 2x_{i1}y_i) \\ L'_{b_2} &= \sum (2x_{i2}^2 b_2 + 2x_{i1}x_{i2}b_1 - 2x_{i2}y_i) \end{cases}$$

示例 — 当 θ 中仅含两个参数b_1, b_2

令 $L'_{b_1} = L'_{b_2} = 0$

$$\begin{cases} (\sum x_{i1}^2)b_1 + (\sum x_{i1}x_{i2})b_2 = \sum x_{i1}y_i \\ (\sum x_{i1}x_{i2})b_1 + (\sum x_{i2}^2)b_2 = \sum x_{i2}y_i \end{cases}$$

示例 — 当 θ 中仅含两个参数b_1, b_2

令 $L'_{b_1} = L'_{b_2} = 0$

$$\begin{cases} (\sum x_{i1}^2)b_1 + (\sum x_{i1}x_{i2})b_2 = \sum x_{i1}y_i \\ (\sum x_{i1}x_{i2})b_1 + (\sum x_{i2}^2)b_2 = \sum x_{i2}y_i \end{cases}$$

显而易见，以上即方程组

$$\begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = (\mathrm{X'X})\mathrm{b} = \mathrm{X'y}$$

线性模型普通最小二乘估计

该方程组解

$$\hat{\theta} = \hat{\mathrm{b}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

目录

育种和育种值

从加性遗传模型理解育种值

- 随机数
- 模型/建模

线性模型普通最小二乘估计和育种值估计

- 方程组
- 普通最小二乘 OLS
- 推导
- 模拟示例

模拟

- 一个最小二乘估计育种值的模拟
- 见 day-1-least-square.ipynb