Introduction to Random Regression (RR) for Test-Day Models

Napo Vargas Jurado

MiX99 course: test-day models and single step genomic prediction

COURSE DAY 1, April 10th, 2025

Linear regression - overview

Model and assumptions

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

- y_{ij}: response variable
- x_{ij}: covariable (continuous)
- β_0: intercept (constant)
- β_1: slope (constant)
- ϵ_{ij}: residual (not necessarily Normal)
- $\hat{\beta}_0$ and $\hat{\beta}_1$ found by minimizing sums of squares (least-squares)

Luke

Example

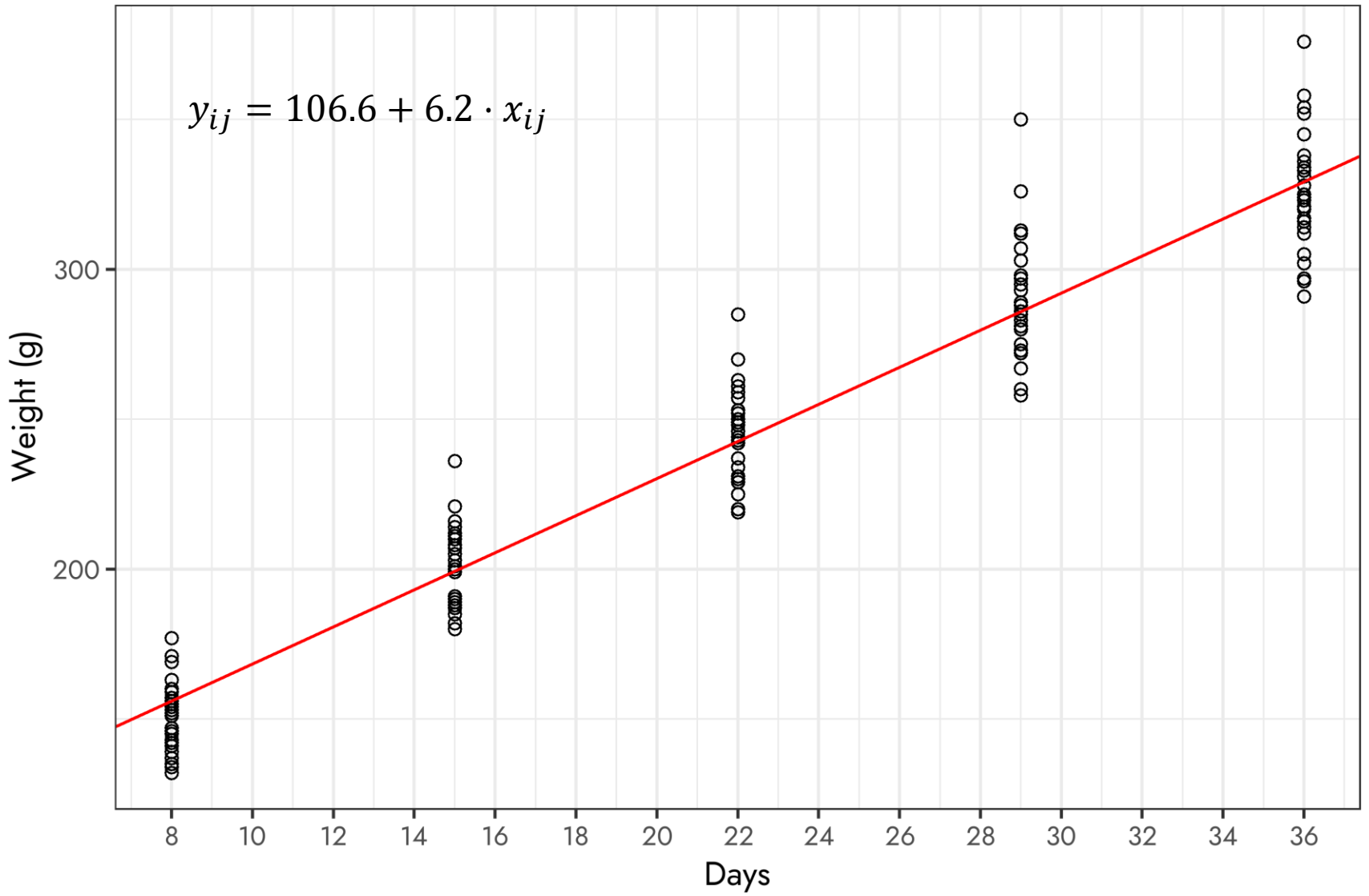
- Weight of rats (Gelfand et al., 1990)

Rat	Weight	Day
1	151	8
1	199	15
⋮	⋮	⋮
30	244	22
30	286	29
30	324	36

- y_{ij}: weight of rat i at day j
- x_{ij}: day of recording

Luke

Example (cont.)



Random regression - overview

Statistical model

- Add a random intercept (b_{0_i}) and random slope (b_{1_i}) for each individual
  - These are deviations from fixed regression
- Random intercept and slope are associated with a distribution

$$\begin{bmatrix} b_{0_i} \\ b_{1_i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right)$$

- Modified regression equation:

$$y_{ij} = (\beta_0 + b_{0_i}) + (\beta_1 + b_{1_i})x_{ij} + \epsilon_{ij}$$

Overview

- Much more flexible model
  - A regression line for each individual
- Additional parameters to estimate (including covariance matrix)
  - REML (or Gibbs sampler)
- For the rat weight example:
  - Fixed intercept and slope ($\hat{\beta}_0$,$\hat{\beta}_1$)
  - 30 random intercepts and slopes ($\hat{b}_{0_i}$, $\hat{b}_{1_i}$)
  - Covariance matrix of random intercept and slopes ($\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$,$\hat{\sigma}_{01}$)
  - 65 parameters in total

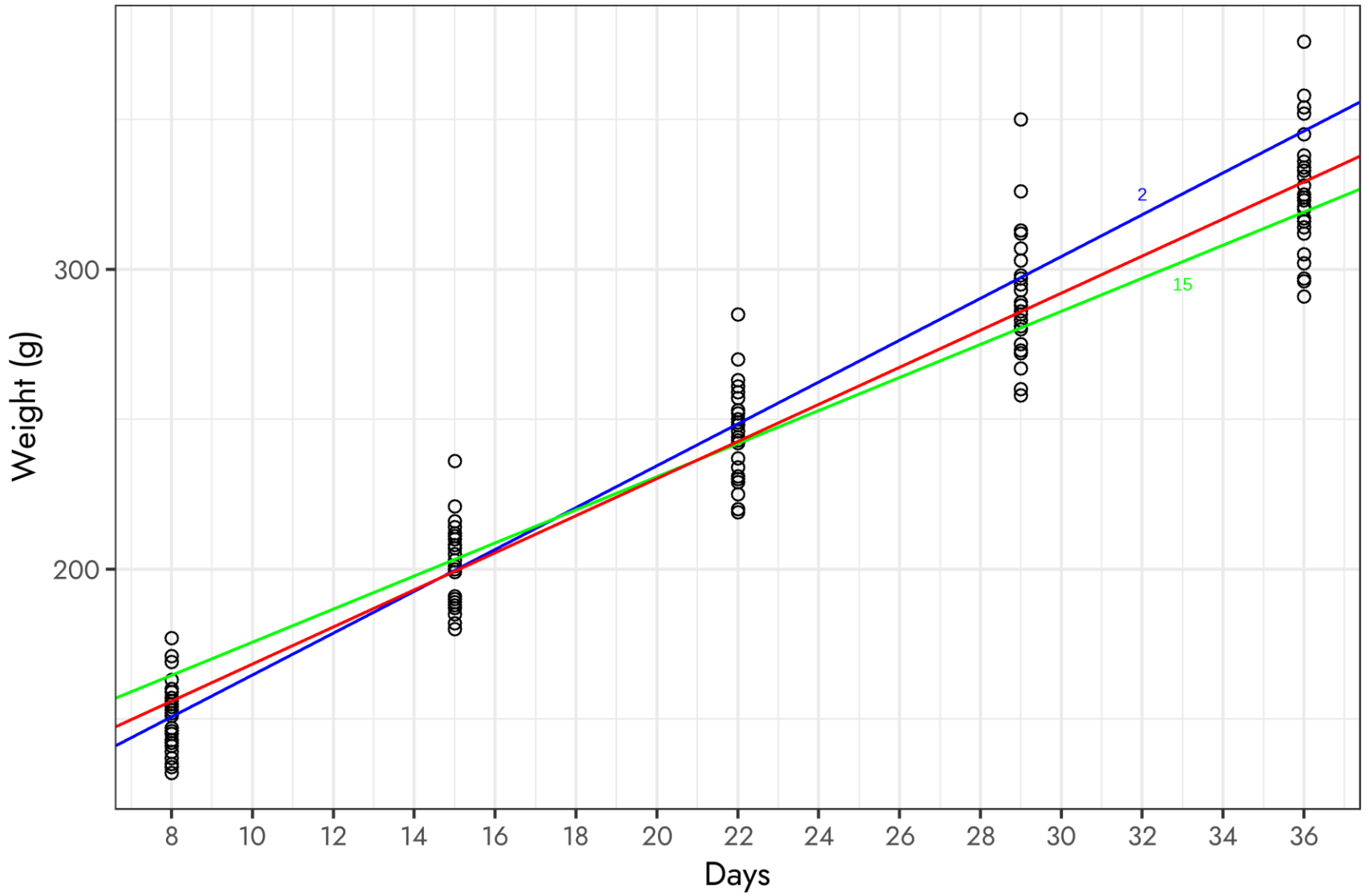
Luke

Example: rats 2 and 15

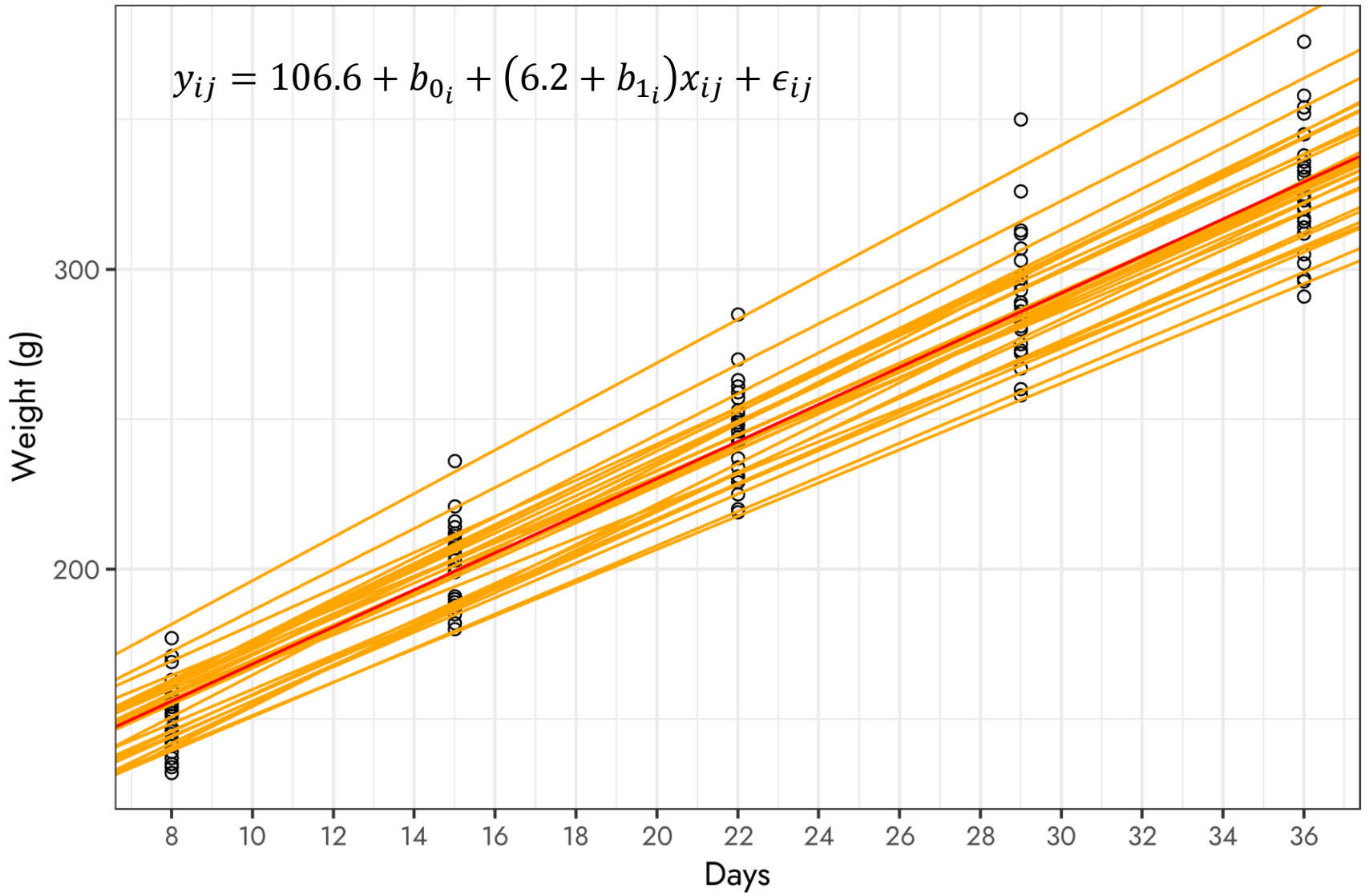
- From fixed regression:
  - $\hat{\beta}_0 = 106.6$
  - $\hat{\beta}_0 = 6.2$
- From random regression
  - $b_{0_2} = -11.6,\ b_{1_2} = 0.8$
  - $b_{0_9} = 14.1, b_{1_9} = -0.7$
- New regression equations:
  - $y_{2j} = (106.6 - 11.6) + (6.2 + 0.8)x_j$
  - $y_{15j} = (106.6 + 14.1) + (6.2 - 0.7)x_j$

Luke

Fixed and random regressions



All rats



Random regression models in animal genetics

Overview

- First introduced to animal breeding and genetics by Henderson (1982):
  - "Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions"
- Popularized by Schaeffer et al., in 1990s for analysis of test-day records
- Lidauer and Strandén (1999) used PCG for solving large RRM models
- RR has become the standard method for analysis of daily milk yield and milk quality (including SCS) records
  - Legendre polynomials
  - Covariance functions (rank reduction)

Luke

Basic model

- Recall the simple random regression model:

$$y_{ij} = \beta_0 + b_{0_i} + \left(\beta_1 + b_{1_i}\right)x_{ij} + \epsilon_{ij}$$

- Replace plain x_{ij} with a function of x_{ij} (e.g., Legendre polynomial, splines):

$$y_{ij} = \beta_0 + \gamma_{0_i} + \left(\beta_1 + \gamma_{1_i}\right)\phi_{ij} + \epsilon_{ij}$$

- Estimate them as before (REML)

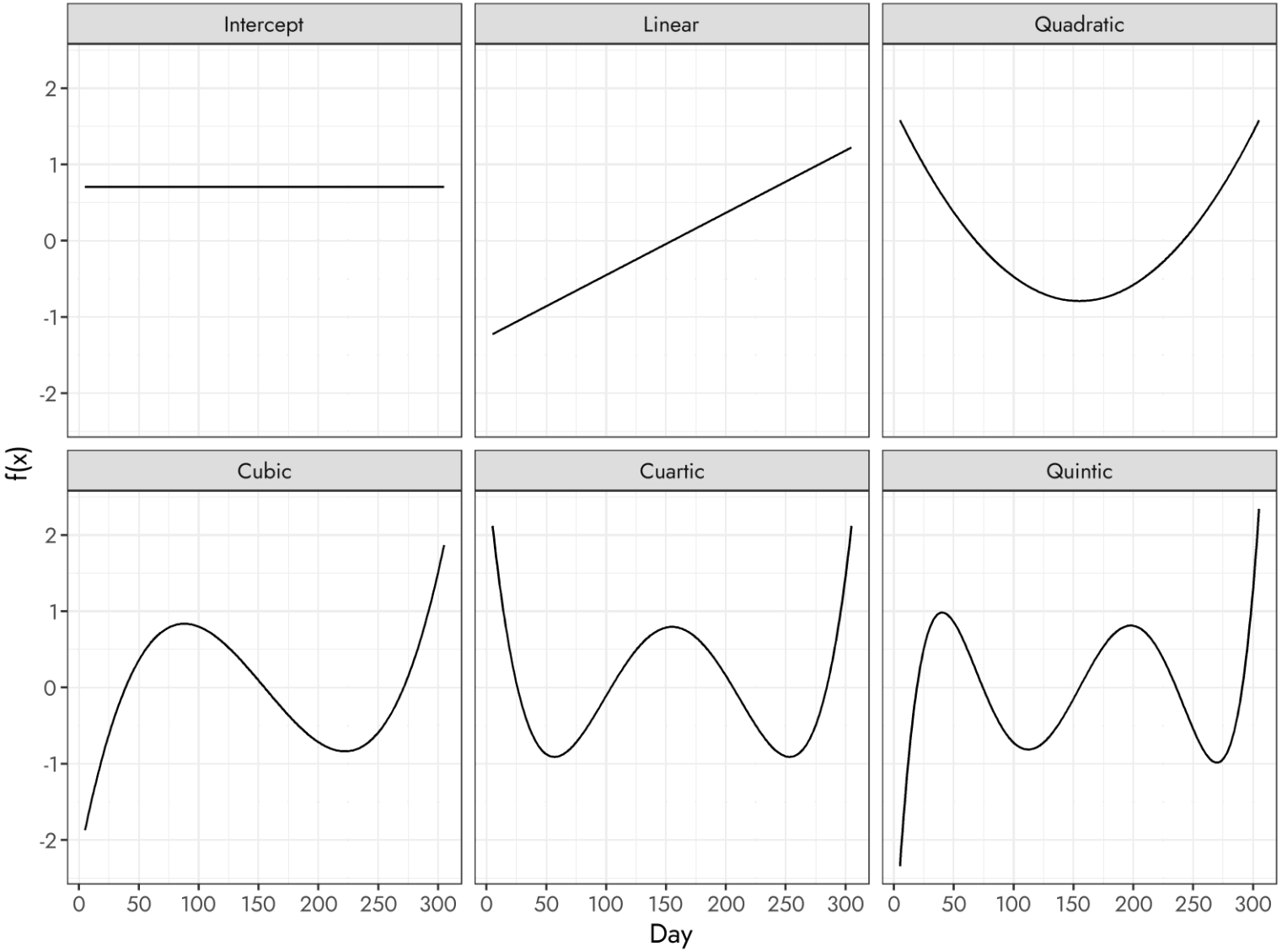
Luke

Legendre polynomials (LP)

- Introduced by Kirkpatrick et al. (1990) to model longitudinal data
  - Continuous functions defined on the interval [-1, 1]
- For 2nd order polynomial:
  - $\phi_0(t) = \sqrt{\frac{1}{2}}, \phi_1(t) = \sqrt{\frac{3}{2}}t, \phi_2(t) = \sqrt{\frac{5}{2}}\left(\frac{3}{2}t^2 - \sqrt{\frac{1}{2}}\right)$
  - $t = 2\left(\frac{d_i - d_{min}}{d_{max} - d_{min}}\right) - 1$
  - d_i is i^{th} day of lactation, d_{min} is minimum of DIM, d_{max} is the maximum of DIM.

Luke

Legendre polynomials



Wilmink function

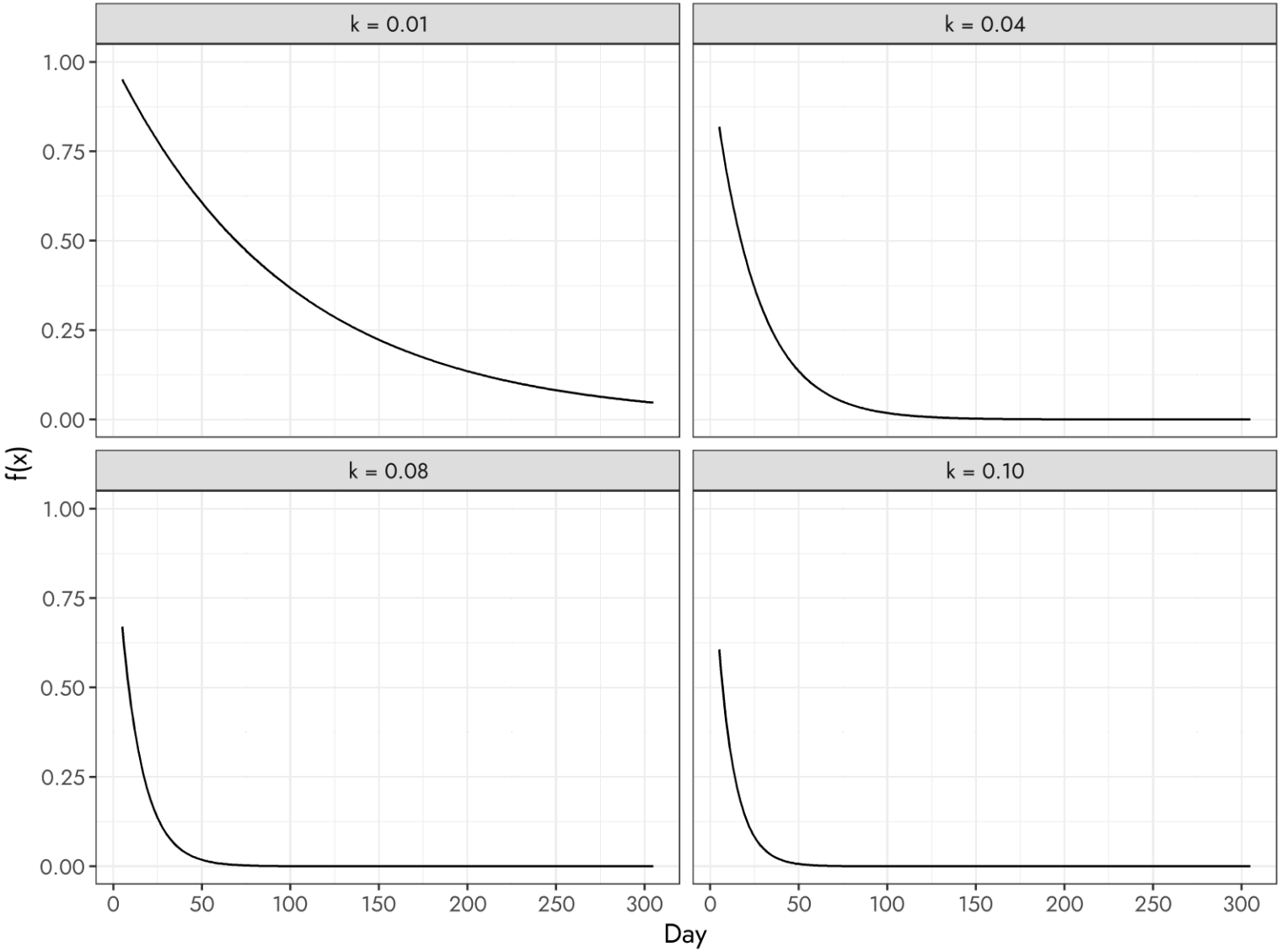
- Not a Legendre polynomial but an exponential function

$$f(t) = e^{-kt}$$

- t is DIM
- k is a constant rate of decay
  - Smaller k → slower decay
  - Larger k → faster decay
- More impact on first days of lactation
  - Close to 0 by the middle of lactation

Luke

Wilmink function



RRM for 2nd order polynomial

$$y_{ij} = \beta_0\phi_0 + \beta_1\phi_{1_i} + \beta_2\phi_{2_i} + \gamma_{0_i}\phi_0 + \gamma_{1_i}\phi_{1_i} + \gamma_{2_i}\phi_{2_i} + \epsilon_{ij}$$

- Fixed regression
- Random regression
- Can be alternatively shown as

$$y_{ij} = \underbrace{\sum_{j=0}^{2} \beta_{ij}\phi_{ij}}_{\text{Fixed}} + \underbrace{\sum_{j=0}^{2} \gamma_{ij}\phi_{ij}}_{\text{Random}} + \epsilon_{ij}$$

Random regression for test-day models

Random regression for test-day records

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{Zc} + \mathbf{e}$$

- **y**: the vector of test-day records
- **Xβ**: fixed effects (including fixed regression curve)
- **Za**: random genetic component (Z contains Legendre polynomials)
- **Zc**: random permanent environmental component
- **e**: vector of residuals

RRM in more detail

$$y_{ik} = \sum_{j=1}^{p} \phi_{ijk}\beta_{ijk} + \sum_{j=1}^{d} \phi_{ijk}\alpha_{ijk} + \sum_{j=1}^{d} \phi_{ijk}\gamma_{ijk} + e_{ik}$$

- y_{ik}: test-day record of cow *i* at day *k (k = 5,…,365)*
- ϕ_{ijk}: value of Legendre Polynomial (**LP**)
- α_{ijk}: RR coefficients for permanent environmental effect
- γ_{ijk}: RR coefficients for genetic effect
- e_{ik}: residual
- Order of LP for fixed regression may differ than that of random regression (*d = p = 2*)

Model assumptions (for 2nd order LP)

$$\begin{bmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}, \mathbf{K}_\alpha = \begin{bmatrix}\sigma^2_{\alpha_0} & \sigma_{\alpha_{01}} & \sigma_{\alpha_{02}}\\ & \sigma^2_{\alpha_1} & \sigma_{\alpha_{12}}\\ \text{Sym.} & & \sigma^2_{\alpha_2}\end{bmatrix}\right)$$

$$\begin{bmatrix}\gamma_0\\ \gamma_1\\ \gamma_2\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}, \mathbf{K}_\gamma = \begin{bmatrix}\sigma^2_{\gamma_0} & \sigma_{\gamma_{01}} & \sigma_{\gamma_{02}}\\ & \sigma^2_{\gamma_1} & \sigma_{\gamma_{12}}\\ \text{Sym.} & & \sigma^2_{\gamma_2}\end{bmatrix}\right)$$

$$e \sim N(0, \sigma^2_e)$$

Model assumptions (cont.)

$$\boldsymbol{\alpha} \sim N(0, \mathbf{K}_{\alpha} \otimes \mathbf{I})$$
$$\boldsymbol{\gamma} \sim N(0, \mathbf{K}_{\gamma} \otimes \mathbf{A})$$

- **A**: numerator relationship matrix (can be replaced by genomic info)
- $\mathbf{I}_d$: Identity matrix.
- Gets more complex (notation) for multiple-trait models
- Usually interested in:
  - Functions of $\boldsymbol{\gamma}$ for genetic evaluation (EBV)
  - Functions of $\mathbf{K}_{\alpha}$, $\mathbf{K}_{\gamma}$, and σ_e^2 for estimation of heritability and genetic correlations (next)

Variances and EBV

- Collecting all LP covariates into a matrix $\mathbf{\Phi}$ (305×3)
- Variability for all DIM can be obtained by:

$$\mathbf{V}_C = \mathbf{\Phi}\mathbf{K}_\alpha\mathbf{\Phi}'$$
$$\mathbf{V}_G = \mathbf{\Phi}\mathbf{K}_\gamma\mathbf{\Phi}'$$
$$\mathbf{V}_P = \mathbf{V}_C + \mathbf{V}_G + \mathbf{R}$$

- Heritability for day i:

$$h_i^2 = \frac{V_{G_{ii}}}{V_{C_{ii}} + V_{G_{ii}} + R_{ii}}$$

- EBV for day i:

$$EBV_i = \hat{\gamma}_0\phi_{0_i} + \hat{\gamma}_1\phi_{1_i} + \hat{\gamma}_2\phi_{2_i}$$

Luke

Combined 305-d EBV

- We can also get a combined EBV
- First define the constants (sum of each of the columns of $\mathbf{\Phi}$):

$$c_j = \sum_{i=1}^{305} \Phi_{ij}\,, \mathbf{c} = [c_0, c_1, c_2]$$

- 305-d EBV:

$$EBV_{305} = \hat{\gamma}_0 c_0 + \hat{\gamma}_1 c_1 + \hat{\gamma}_2 c_2$$

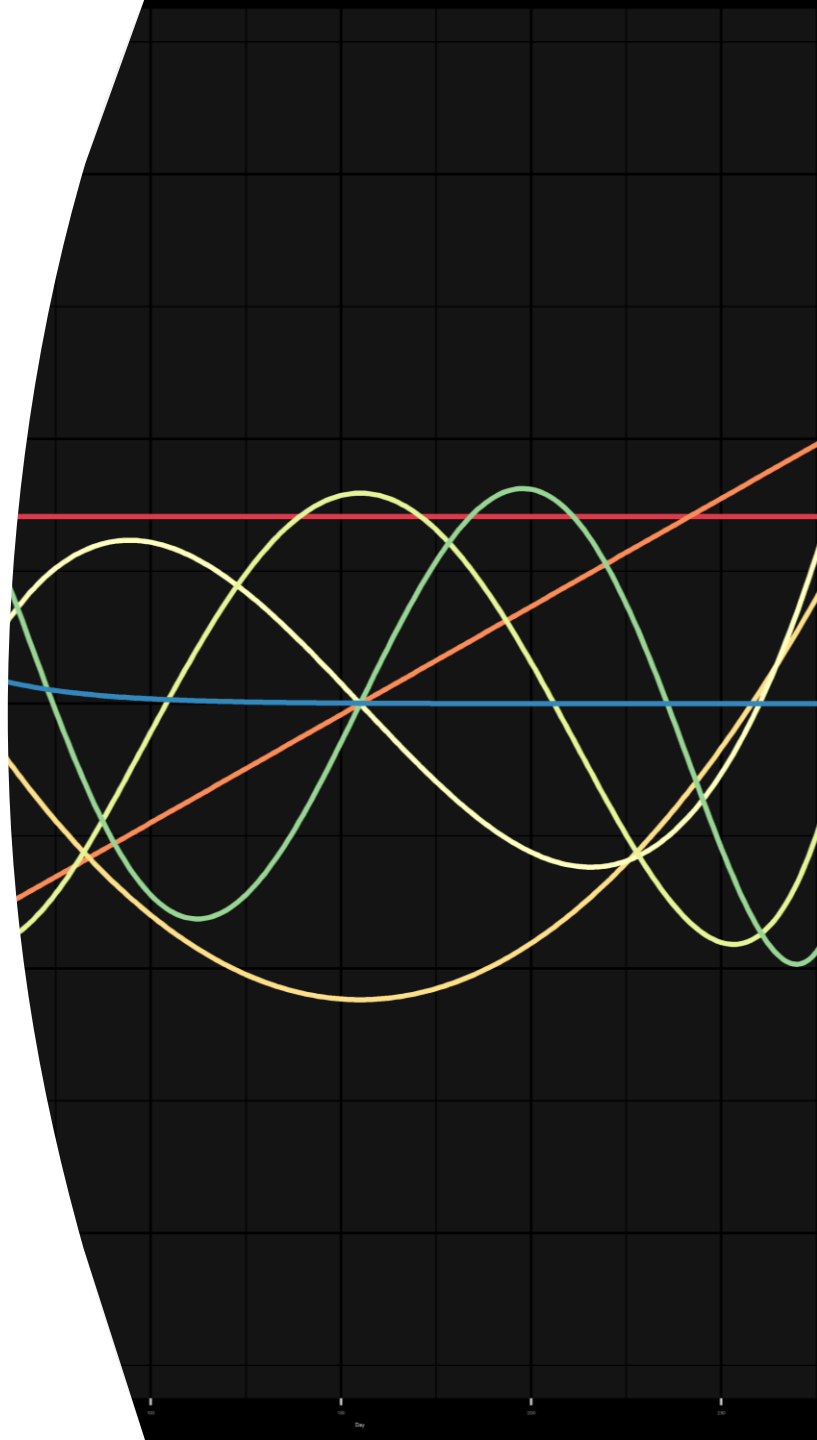
- These can be used to build indexes, rankings, etc.

Luke

Summary

Random regression models

- Flexible strategy for modeling individual variability in test-day data
  - Variability at any given day (or time)
  - Combined EBV
- Standard for test-day records
  - Coupled with Legendre polynomials
- Trade-off
  - Many parameters to estimate (regression coefficients, VC)
    - Genetic and permanent-environmental effect
  - Especially for complex models and many traits
  - Need large number of records



Questions?