Validation des modèles statistiques dans l'analyse de données longitudinales

C. Lopez (Institut de l'Elevage)

Quels modèles ?

Les outils diagnostics

(modèles à effets fixes)

$$Y = X\beta + \varepsilon \quad \varepsilon \square N\left(0, \sigma^2 I\right)$$

$$\text{var}\left[\hat{\varepsilon}\right] = \sigma^2 \left(I - H\right) = \sigma^2 \left(I - X\left(X'X\right)^{-} X'\right)$$

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}$$

Résidus ‘studentisés’

$$r^*_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2}}$$

Résidus de Pearson
(résidus de Cholesky)

t_i Résidus ‘studentisés’ externe

Mesure de l'influence des observations

analyse avec et sans l'observation i

et on compare…

Influence globale

$$PRESS = \sum_i \left(y_i - \hat{y}_{(-i)} \right)^2$$

Influence sur l'estimation de β

$$Cook\ D = r_i^2 \frac{1}{rang(X)} \frac{h_{ii}}{(1-h_{ii})}$$

$$DFFITS_i = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

Autres statistiques…

Influence sur la précision des estimations de β

$$\text{cov ratio}(\beta) = \frac{\left|\hat{\text{var}}\left[\hat{\beta}_{(-i)}\right]\right|}{\left|\hat{\text{var}}\left[\hat{\beta}\right]\right|}$$

Les outils diagnostics

(modèles mixtes)

$$Y = X\beta + Z\gamma + \varepsilon$$

$$Var(\gamma) = G \qquad var(\varepsilon) = R$$

$$Y \mid \gamma \sim N\left[X\beta + Z\gamma \; ; \; R \right]$$

$$Y \sim N\left[X\beta \; ; \; ZGZ'+R \right]$$

Deux types de résidus…

$r_c = Y - X\hat{\beta} - Z\hat{\gamma}$ Résidus conditionnels

$r_m = Y - X\hat{\beta}$ Résidus marginaux

sorties différentes pour les deux types de résidus…

La structure de variance-covariance

influe sur les estimations des effets fixes

1- modélisation de la structure de variance-covariance

2- modélisation des effets fixes

Analyse de l'influence des observations plus complexe

Mesure de l'influence des observations

Dans le cas de l'analyse de données longitudinales on s'intéresse à l'influence des individus 'clusters' plutôt qu'à l'influence d'observations isolées

Mêmes statistiques d'influence que

dans le cas du modèle à effets fixes (1)

Appliquées aux paramètres β

et aux paramètres de variance-covariance G et R

(1) + « Likelihood displacement (Cook (1986)) »

ESSAI 'SYSTEME' CRECOM

Comparaison de 2 conduites alimentaires en station expérimentale

2 systèmes fourragers :
25% Maïs (Trait=1 ; n=23) *vs* 50% Maïs (Trait=2 ; n=31)

3 périodes de vêlage :
été (sais=1 ; n=19), hiver (sais=2 ; n=24), printemps (sais=3 ; n=11)

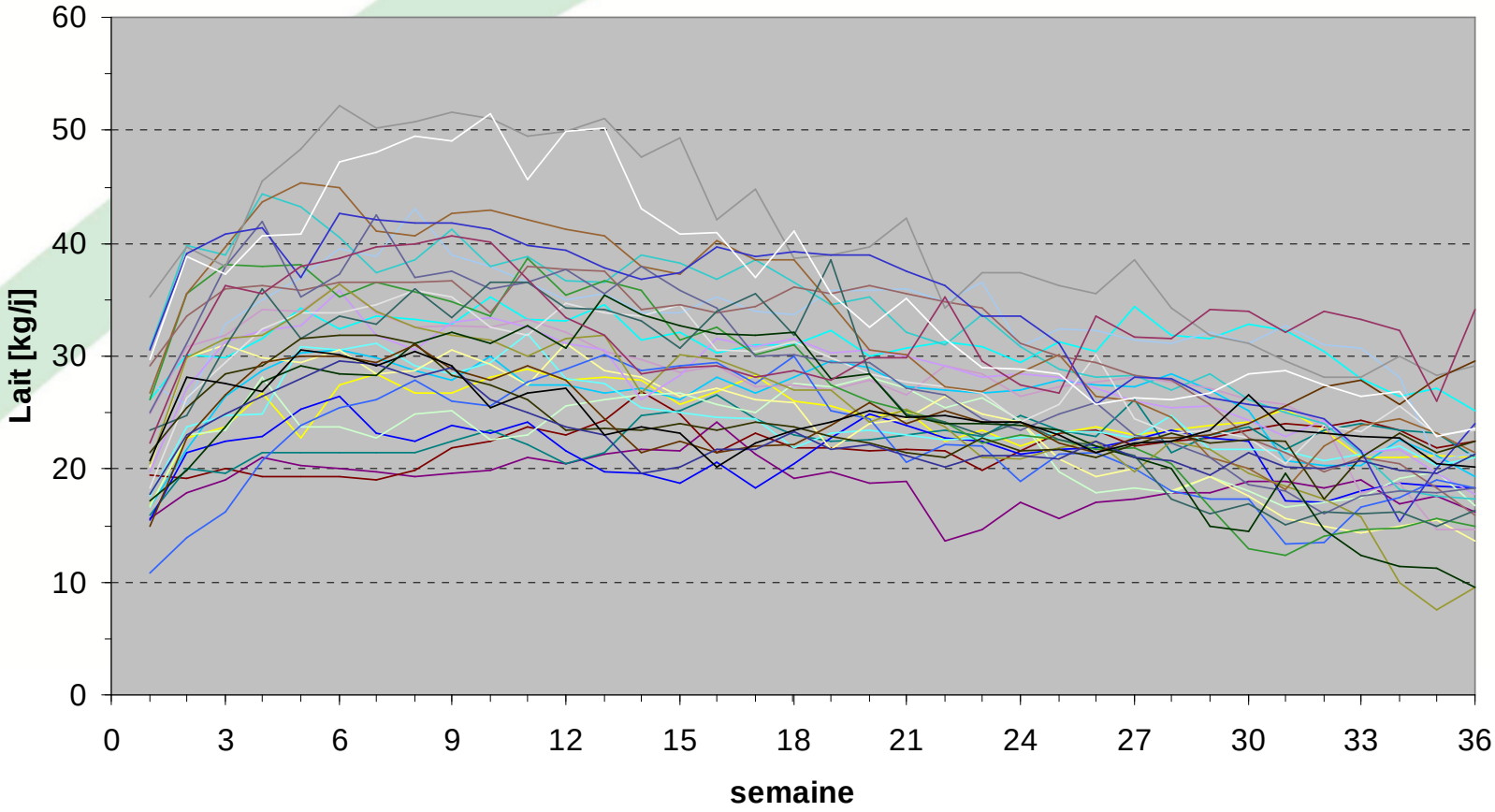
Rang de lactation :
primipares (Parite='P' ; n=23), multipares (Parite='M' ; n=31)

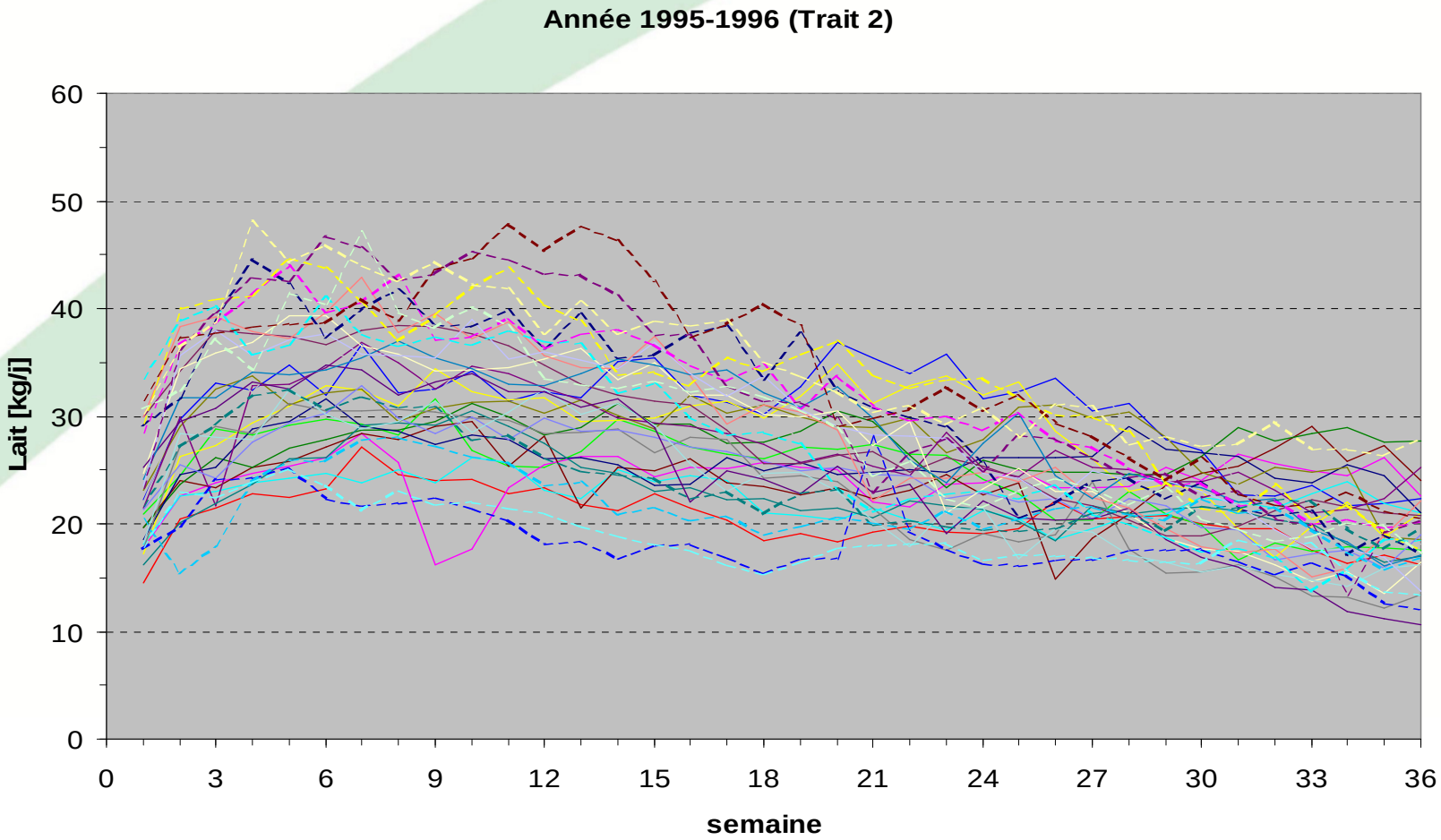
Critère : la production laitière [kg/j]

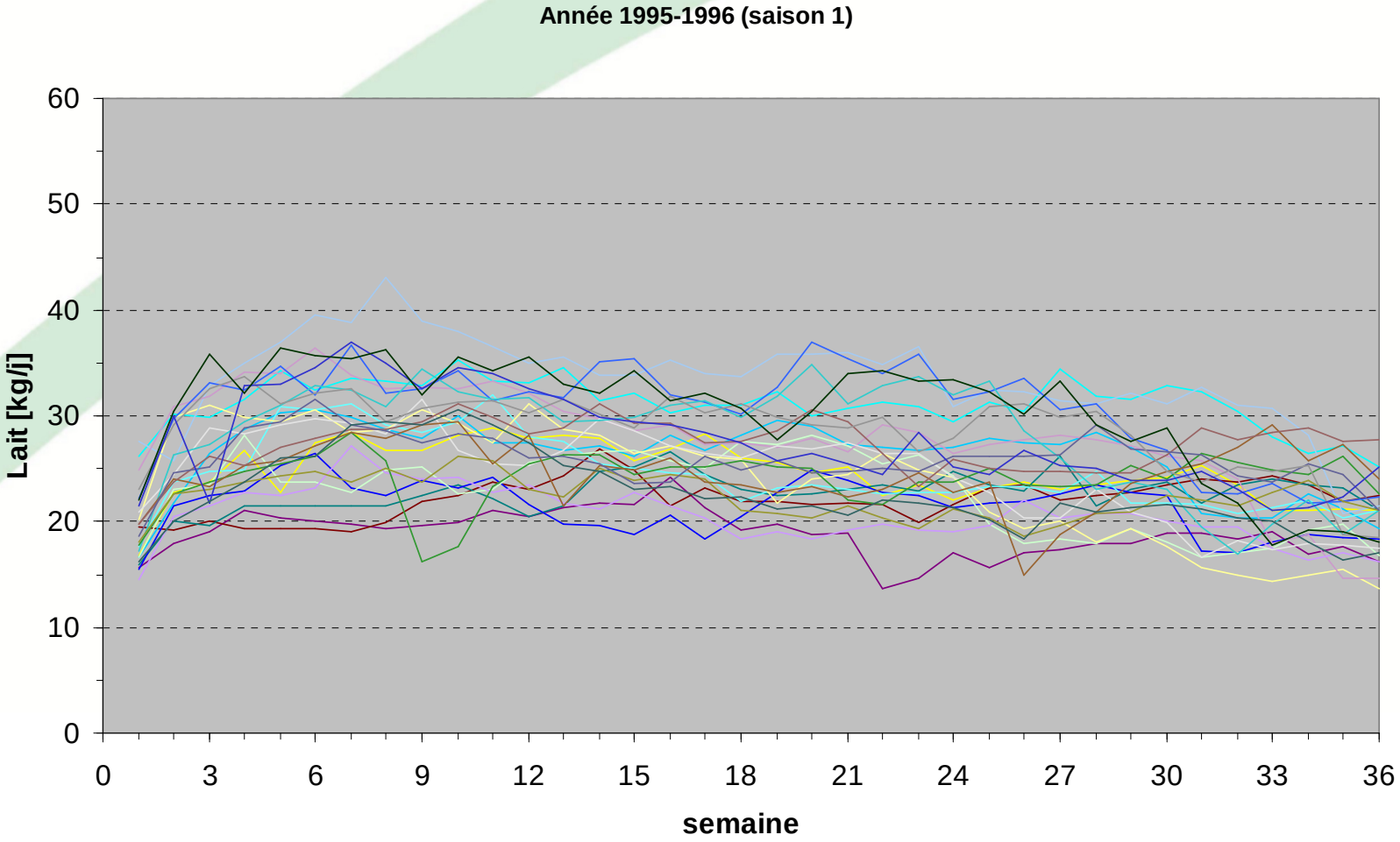
54 vaches Prim'Holstein
36 productions quotidiennes moyennes/semaine

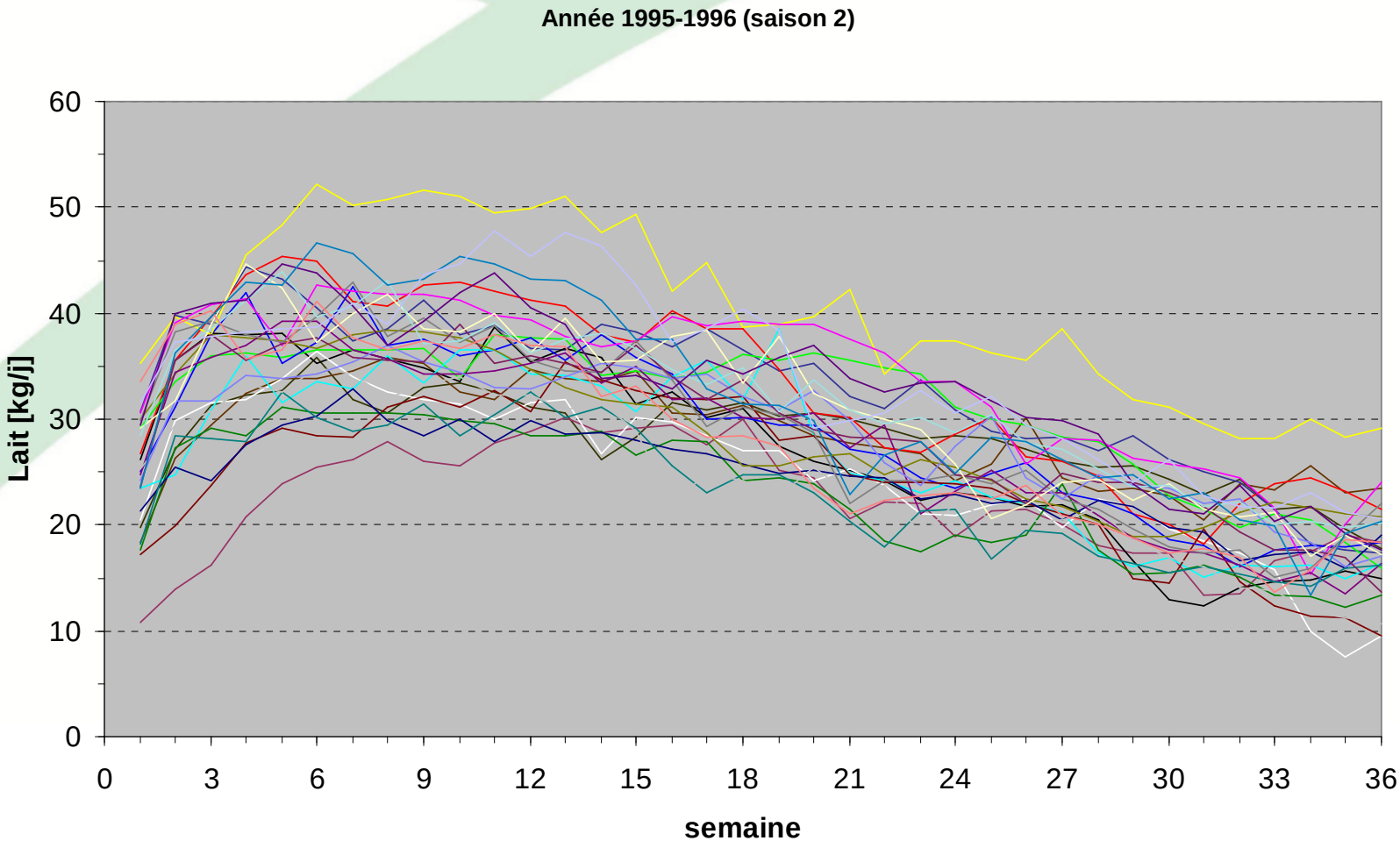
Les données

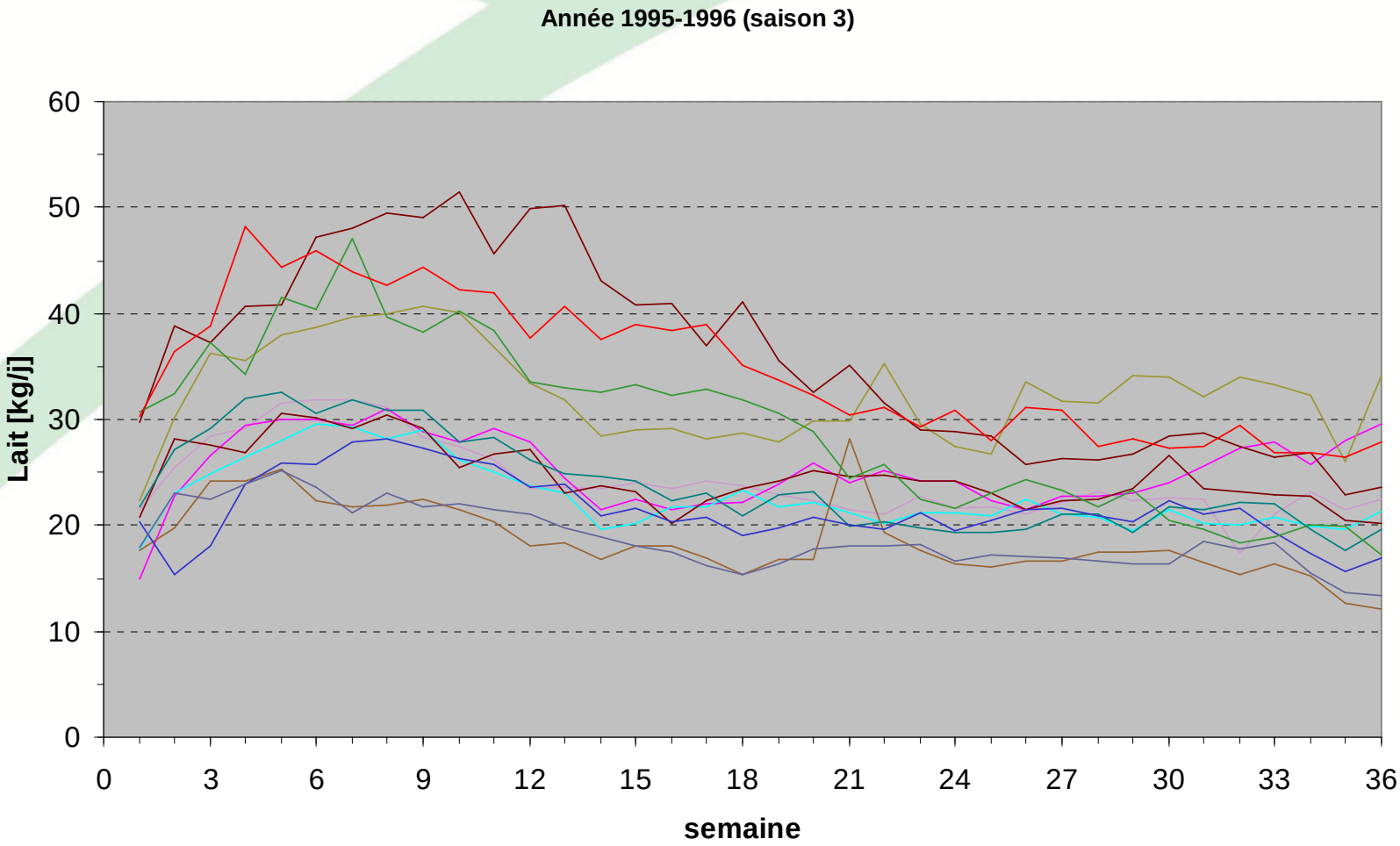
Année 1995-1996 (Trait 1)



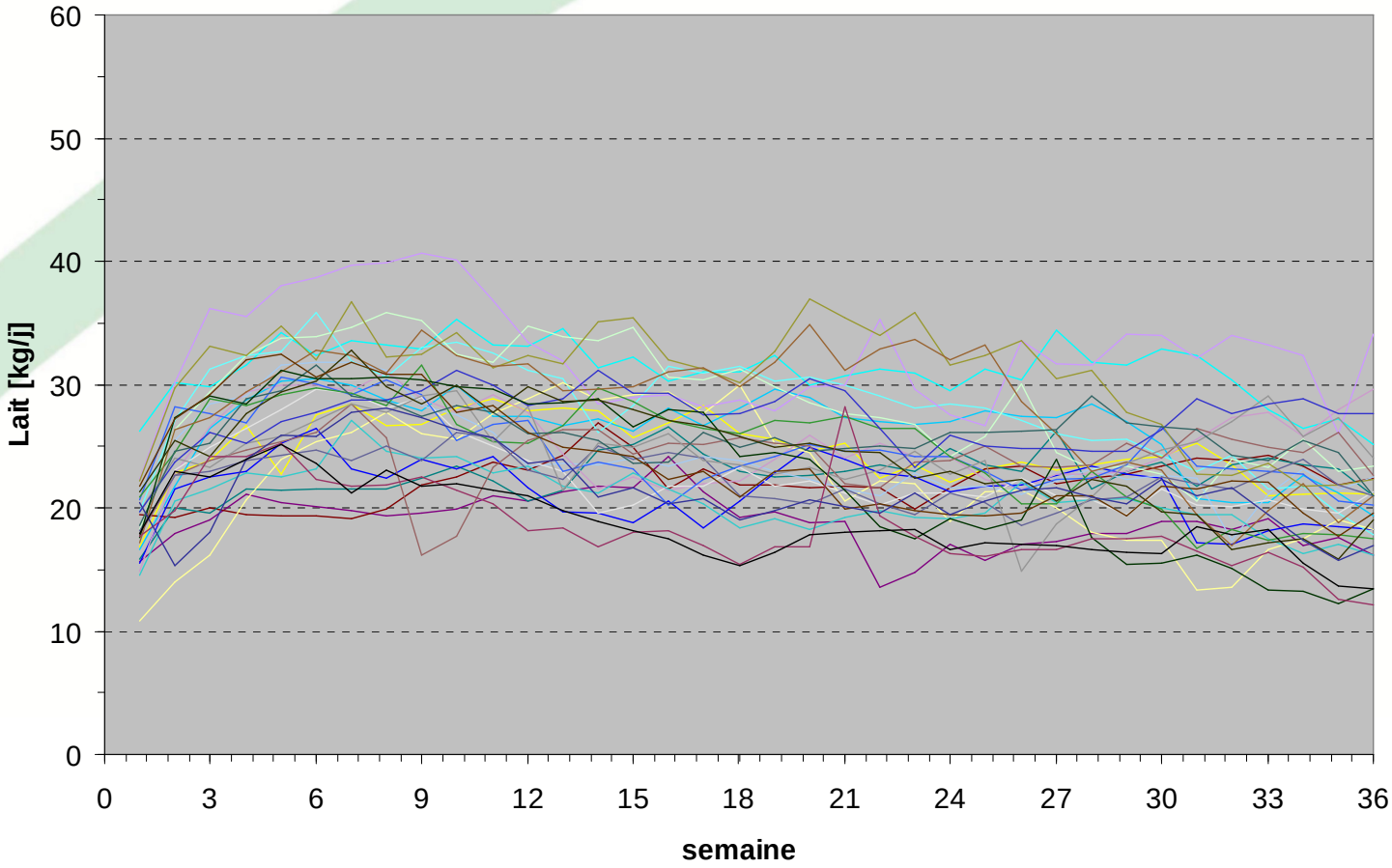

Année 1995-1996 (Trait 2)
Lait [kg/j]
60
50
40
30
20
10
0
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine

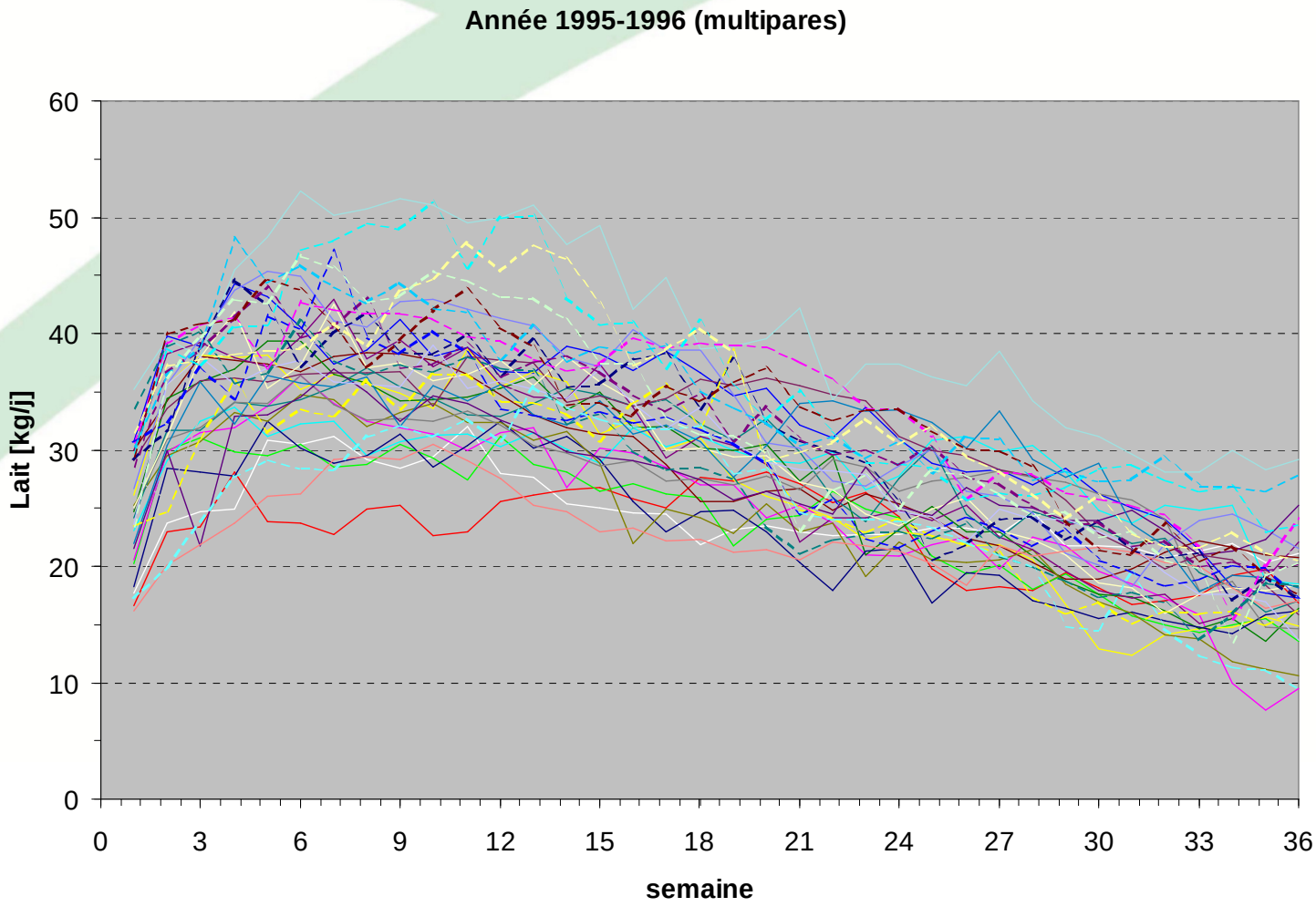

Année 1995-1996 (saison 1)
Lait [kg/j]
60
50
40
30
20
10
0
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine


Année 1995-1996 (saison 2)
Lait [kg/j]
60
50
40
30
20
10
0
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine


Année 1995-1996 (saison 3)
Lait [kg/j]
60
50
40
30
20
10
0
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine

Année 1995-1996 (primipares)




Année 1995-1996 (multipares)
Lait [kg/j]
semaine
0
10
20
30
40
50
60
0
3
6
9
12
15
18
21
24
27
30
33
36

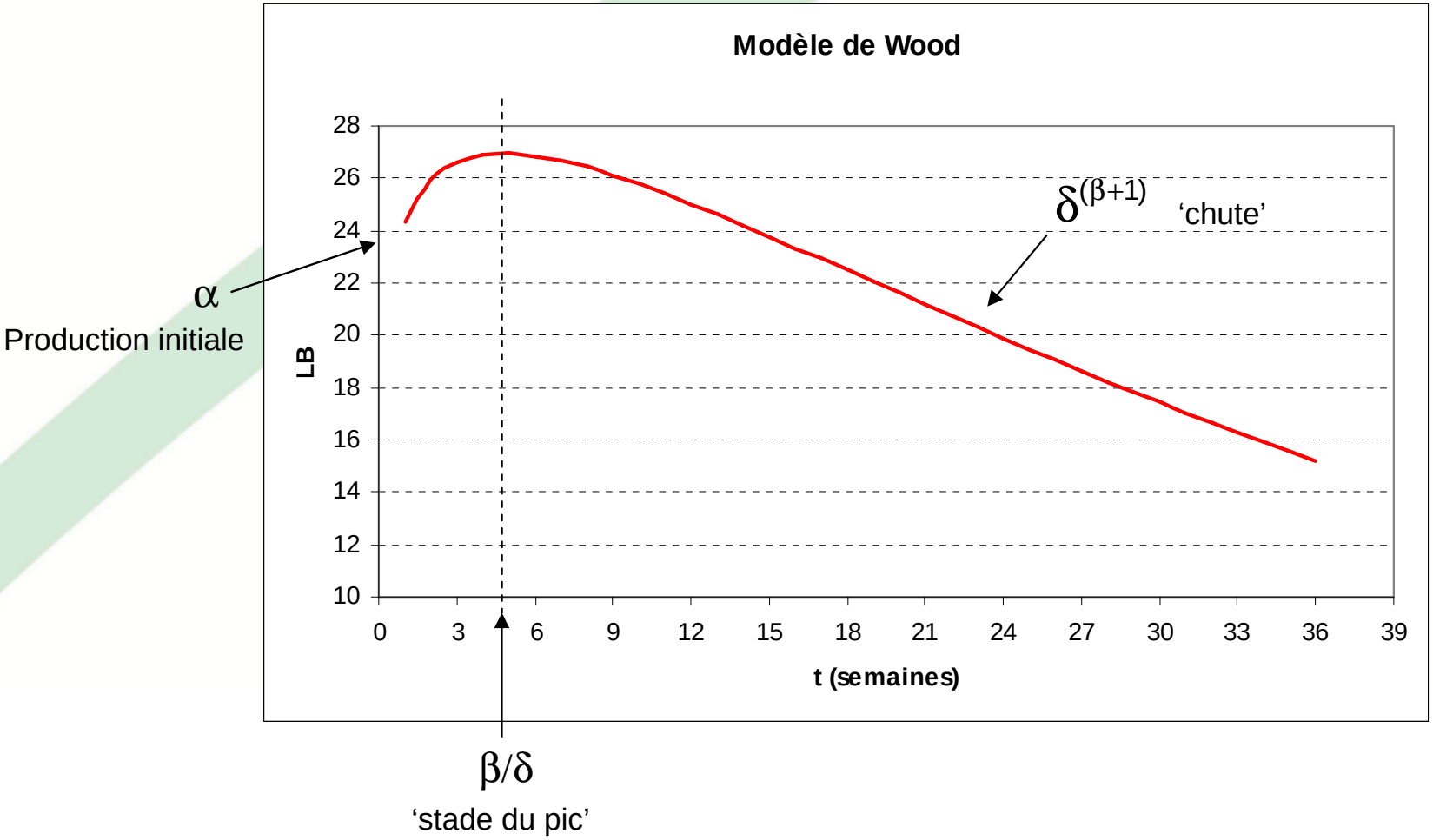
Réponse : profil de mesures hebdomadaires….

…Analyse sur variables résumées de la lactation

Modèle de WOOD (1967)

$$y_t = \alpha * t^{\beta} * e^{-\delta * t}$$

$$0 < \alpha, \beta, \delta < 1$$


Modèle de Wood
LB
28
26
24
22
20
18
16
14
12
10
0
3
6
9
12
15
18
21
24
27
30
33
36
39
t (semaines)
α
Production initiale
δ(β+1)
'chute'
β/δ
'stade du pic'

Le modèle statistique

$$z_t = \mathrm{Log}_e(y_t) = \tilde{\alpha} + \beta * \mathrm{Log}_e(t) - \delta * t$$

$Z_{ijku}(t)$

- i : régime alimentaire
- j : parité (primipares / multipares)
- k : saison(1, 2, 3))
- u : vache (ijk)

Modèle à coefficients aléatoires

$$z_{ijku}(t) = \tilde{\alpha}_{ijk} + \beta_{ijk} * \mathrm{Log}_e(t) - \delta_{ijk} * t \qquad X\beta$$

$$+\tilde{\alpha}_{u(ijk)} + \beta_{u(ijk)} * \mathrm{Log}_e(t) - \delta_{u(ijk)} * t \qquad Z\gamma$$

$$+\varepsilon_{ijku}(t) \qquad \varepsilon$$

La structure de variance-covariance

$$\gamma_{u(ijk)} = \begin{bmatrix} \alpha_{u(ijk)} \\ \beta_{u(ijk)} \\ \delta_{u(ijk)} \end{bmatrix} \quad \text{i.i.d. } N(0 ; G) \quad \forall u(ijk) \qquad G : \text{type} = UN$$

$$\varepsilon_{ijku} = \begin{bmatrix} \varepsilon_{(1)ijku} \\ \varepsilon_{(2)ijku} \\ \dots \\ \\ \varepsilon_{(36)ijku} \end{bmatrix} \quad \text{i.i.d. } N(0 ; R) \quad \forall u(ijk) \qquad R : \text{type} = AR(1)$$

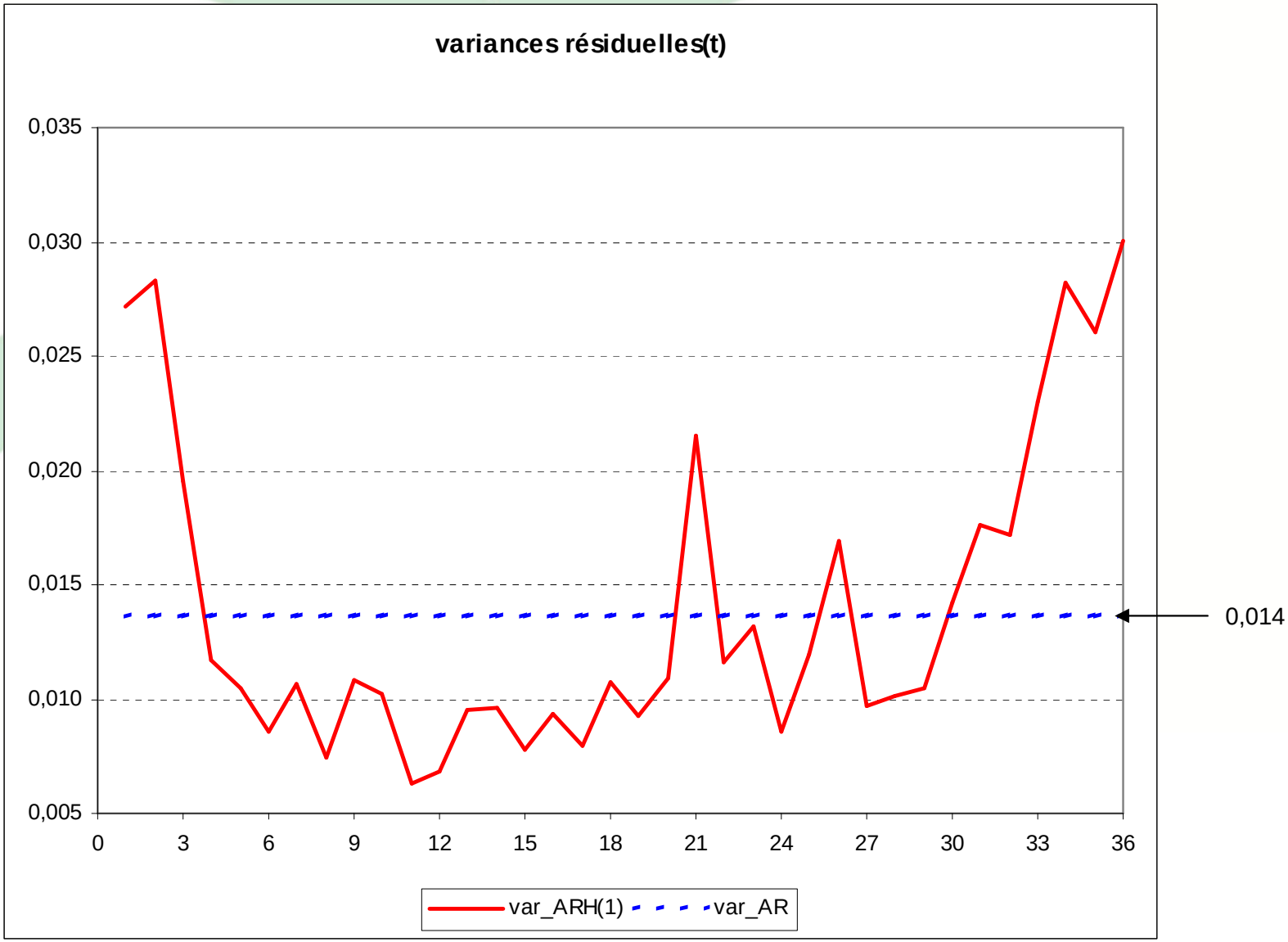
1- Modélisation de la structure de variance-covariance (modèle saturé en Xβ)

2- Modélisation de la partie fixe du modèle

1- Modélisation de la structure de variance-covariance (modèle saturé en Xβ)

Test de différentes structures de matrices R

	2LL_RES	**Khi-Deux**	**ν**	**p**
ARH(1)	4364,5			
AR(1)	4148,4	216,1	35	5,8583E-28
VC	3177,5	970,9	1	3,7988E-213


variances résiduelles(t)
0,035
0,030
0,025
0,020
0,015
0,010
0,005
0
3
6
9
12
15
18
21
24
27
30
33
36
0,014
var_ARH(1)
var_AR

Estimation de la matrice G (modèle saturé en Xβ)

paramètres estimés variance-Covariance					
Parm cov		**Estimation**	**Erreur type**	**Valeur Z**	**Pr Z**
UN(1,1)	σ2(α)	0.01408	0.005883	2.39	0.0084
UN(2,1)	σ(αβ)	-0.00460	0.002805	-1.64	0.1012
UN(2,2)	σ2(β)	0.003026	0.001910	1.58	0.0566
UN(3,1)	σ(αδ)	0.000659	0.000281	2.35	0.0189
UN(3,2)	σ(βδ)	-0.00035	0.000192	-1.83	0.0677
UN(3,3)	σ2(δ)	0.000041	0.000021	1.90	0.0290
AR(1)	ρ	0.7558	0.02360	32.03	<.0001
Residual	ε	0.01359	0.001311	10.37	<.0001

Test de différentes structures de matrices G (simplification du modèle)

Modélisation de la structure de variance-covariance

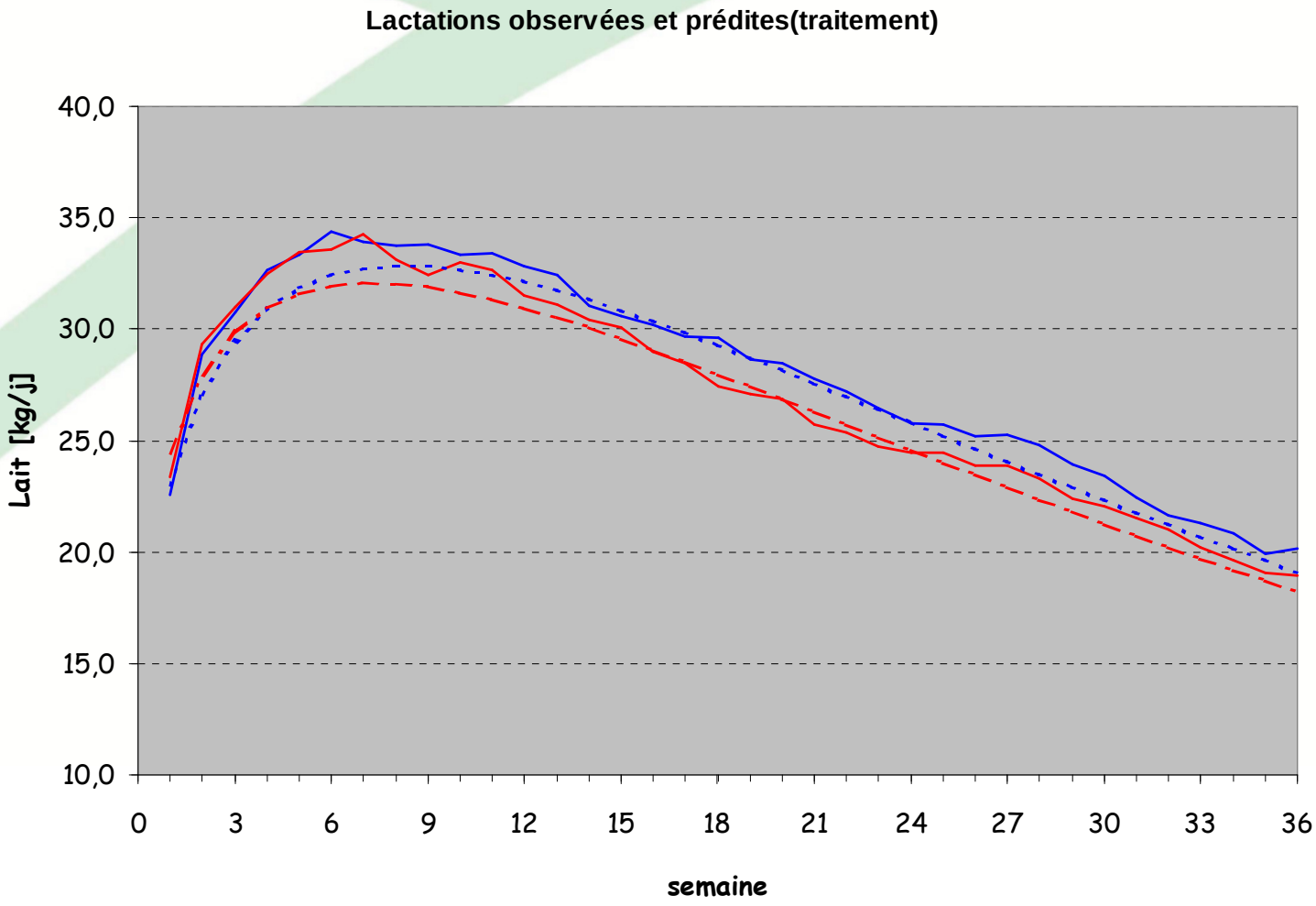
	2LL_RES	**Khi-Deux**	ν	**p**
UN(α,β,δ)	4131,4			
UN(α,δ)	4122,9	8,5	3	0,037
UN(α,β)	4116,2	15,2	3	0,002

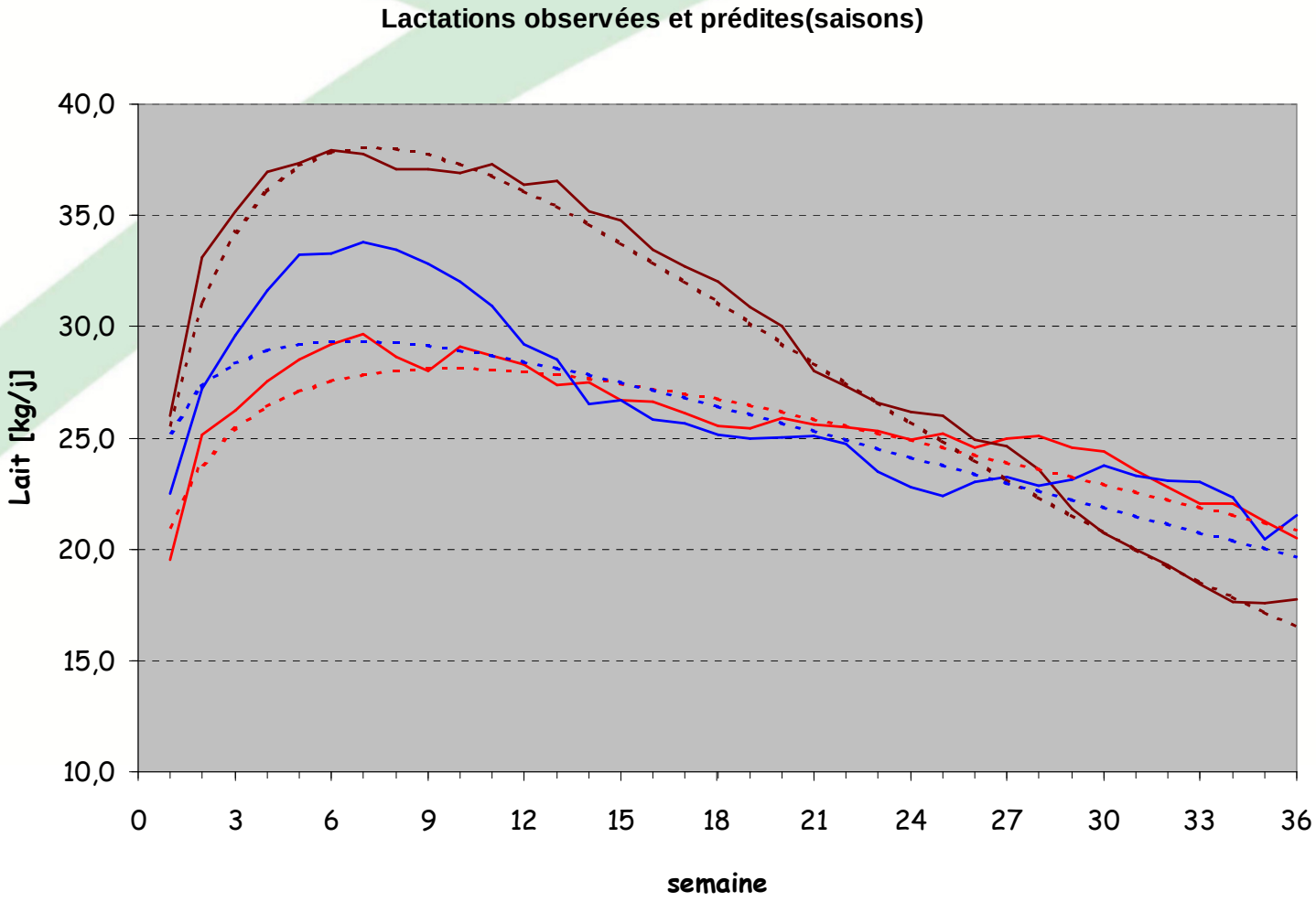
2. Tests des effets fixes

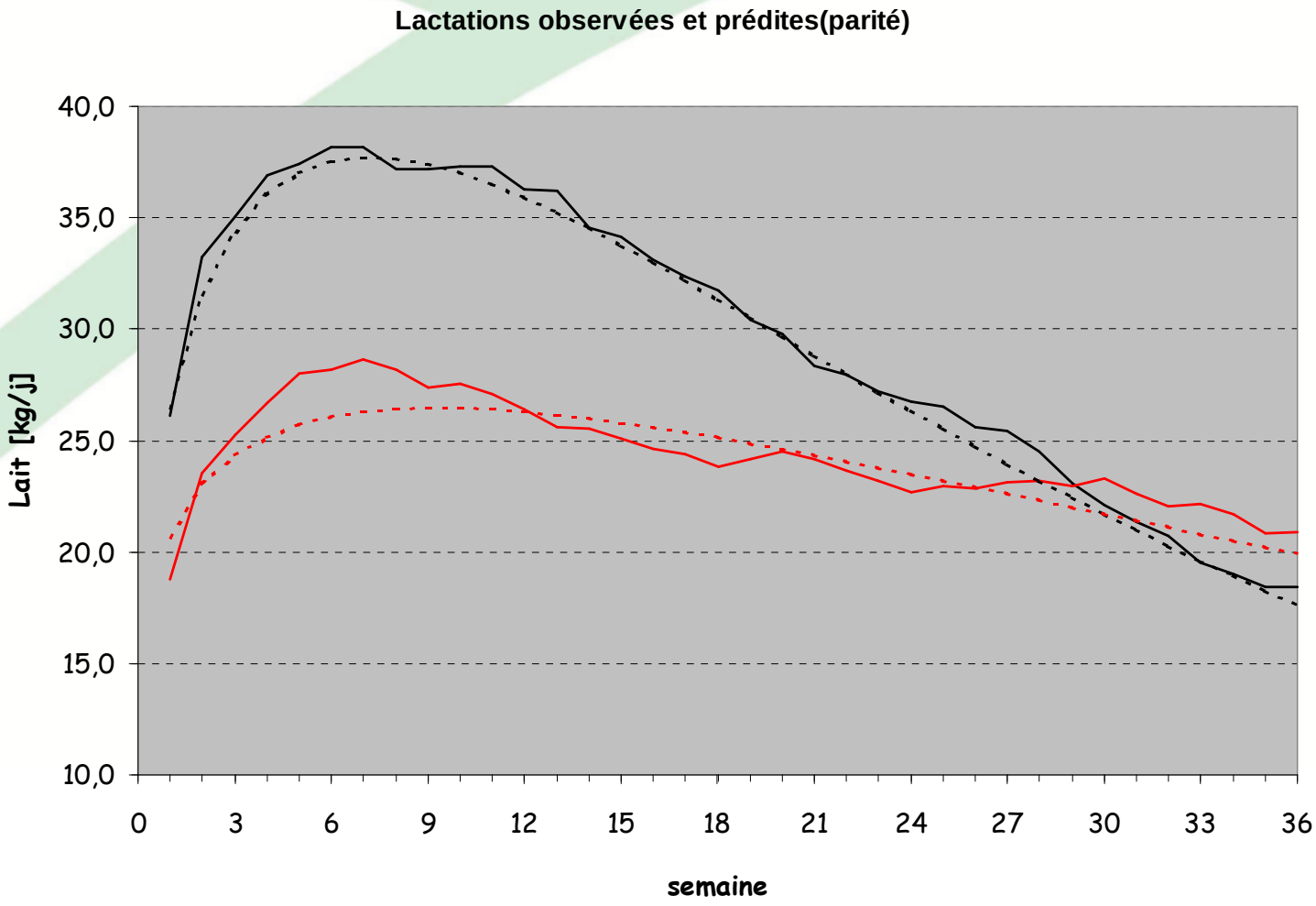
	Type 3 Tests des effets fixes				
	Effet	**Num DF**	**Den DDL**	**F Value**	**Pr > F**
	t	1	44	447.72	<.0001
	L_t	1	44	381.03	<.0001
α	TRAIT	1	44	0.58	0.4511
α	parite	1	44	43.53	**<.0001**
α	SAIS	2	44	6.05	**0.0048**
α	TRAIT*parite	1	44	1.45	0.2344
α	TRAIT*SAIS	2	44	1.29	0.2850
α	parite*SAIS	2	44	6.47	**0.0034**
β	L_t*TRAIT	1	44	3.76	0.0591
β	L_t*parite	1	44	0.49	0.4857
β	L_t*SAIS	2	44	6.26	0.0040
β	L_t*TRAIT*parite	1	44	0.19	0.6691
β	L_t*TRAIT*SAIS	2	44	1.05	0.3583
β	L_t*parite*SAIS	2	44	3.16	0.0521
δ	t*TRAIT	1	44	0.43	0.5167
δ	t*parite	1	44	12.58	**0.0009**
δ	t*SAIS	2	44	10.65	**0.0002**
δ	t*TRAIT*parite	1	44	0.78	0.3822
δ	t*TRAIT*SAIS	2	44	0.06	0.9423
δ	t*parite*SAIS	2	44	1.59	0.2160

Statistiques sur les paramètres-résumés des courbes

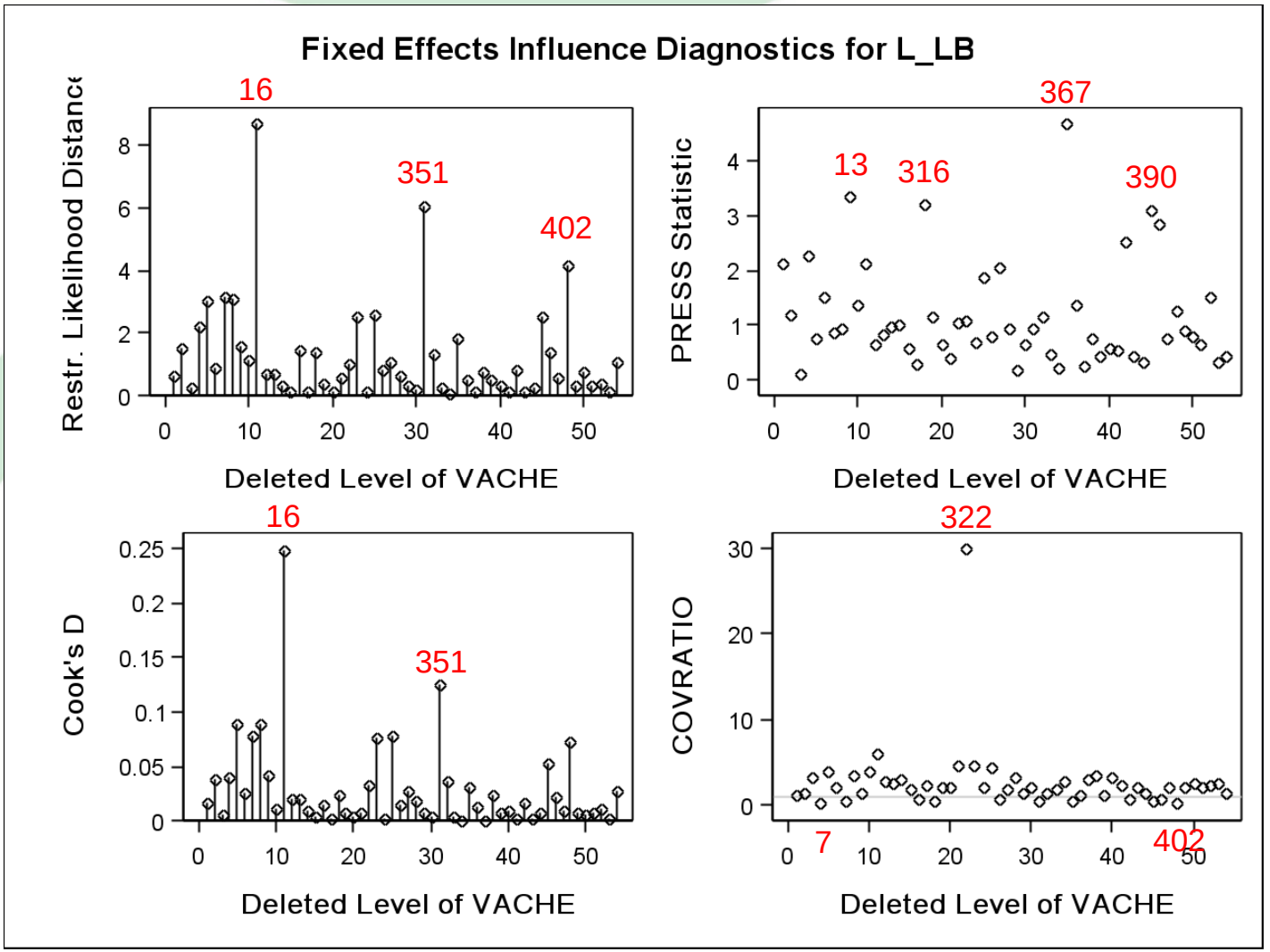
	TRAIT 1		TRAIT 2	
	est	S	est	S
α_est [kg/j]	22,4	0,8	23,5	0,6
stade_Pic [sem]	8,8	0,6	8,0	0,5
chute	1,3%	2,7%	1,4%	2,7%


Lactations observées et prédites(traitement)
40,0
35,0
30,0
25,0
20,0
15,0
10,0
Lait [kg/j]
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine

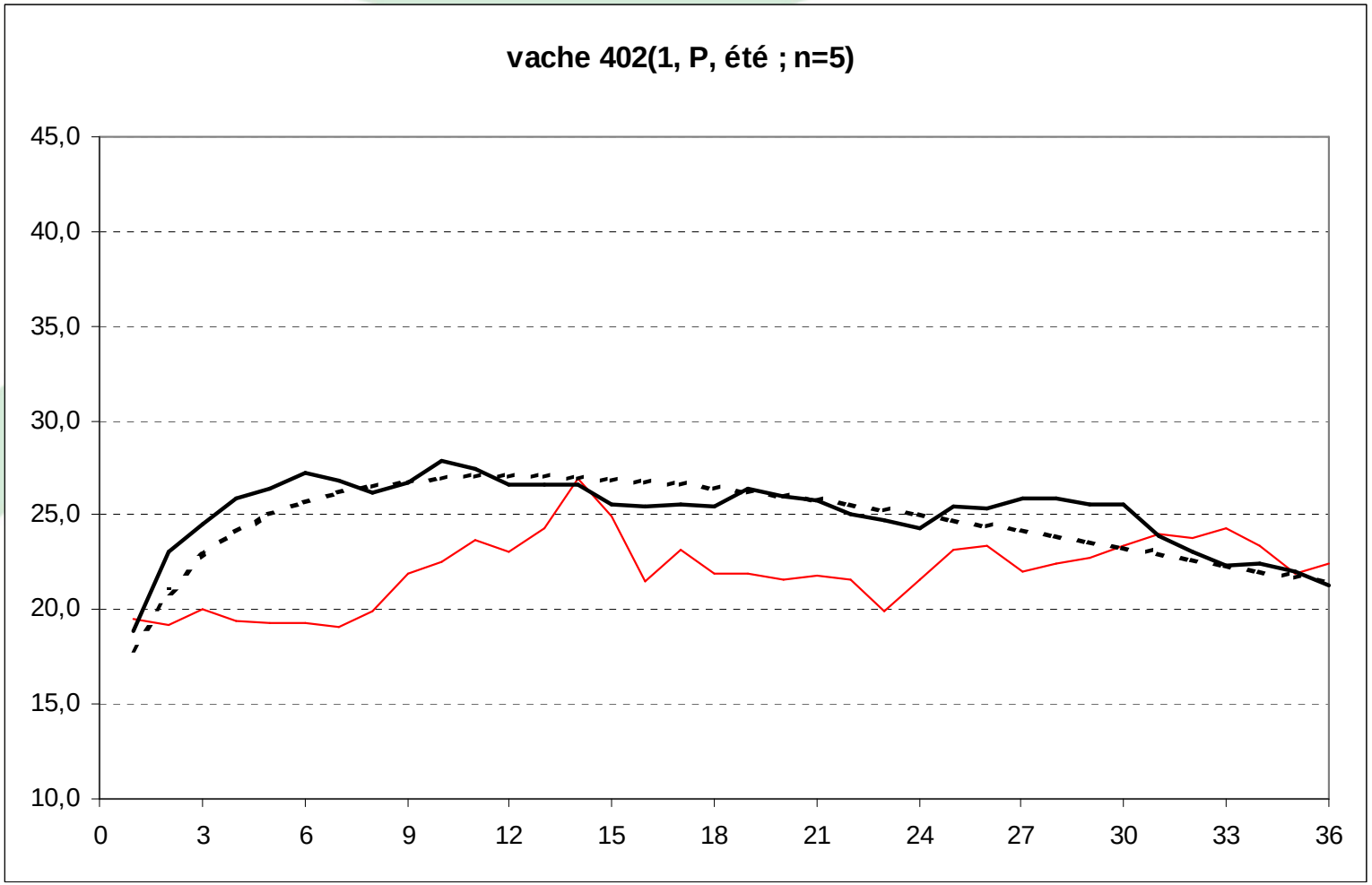

Lactations observées et prédites(saisons)
Lait [kg/j]
40,0
35,0
30,0
25,0
20,0
15,0
10,0
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine


Lactations observées et prédites(parité)
Lait [kg/j]
40,0
35,0
30,0
25,0
20,0
15,0
10,0
0
3
6
9
12
15
18
21
24
27
30
33
36
semaine

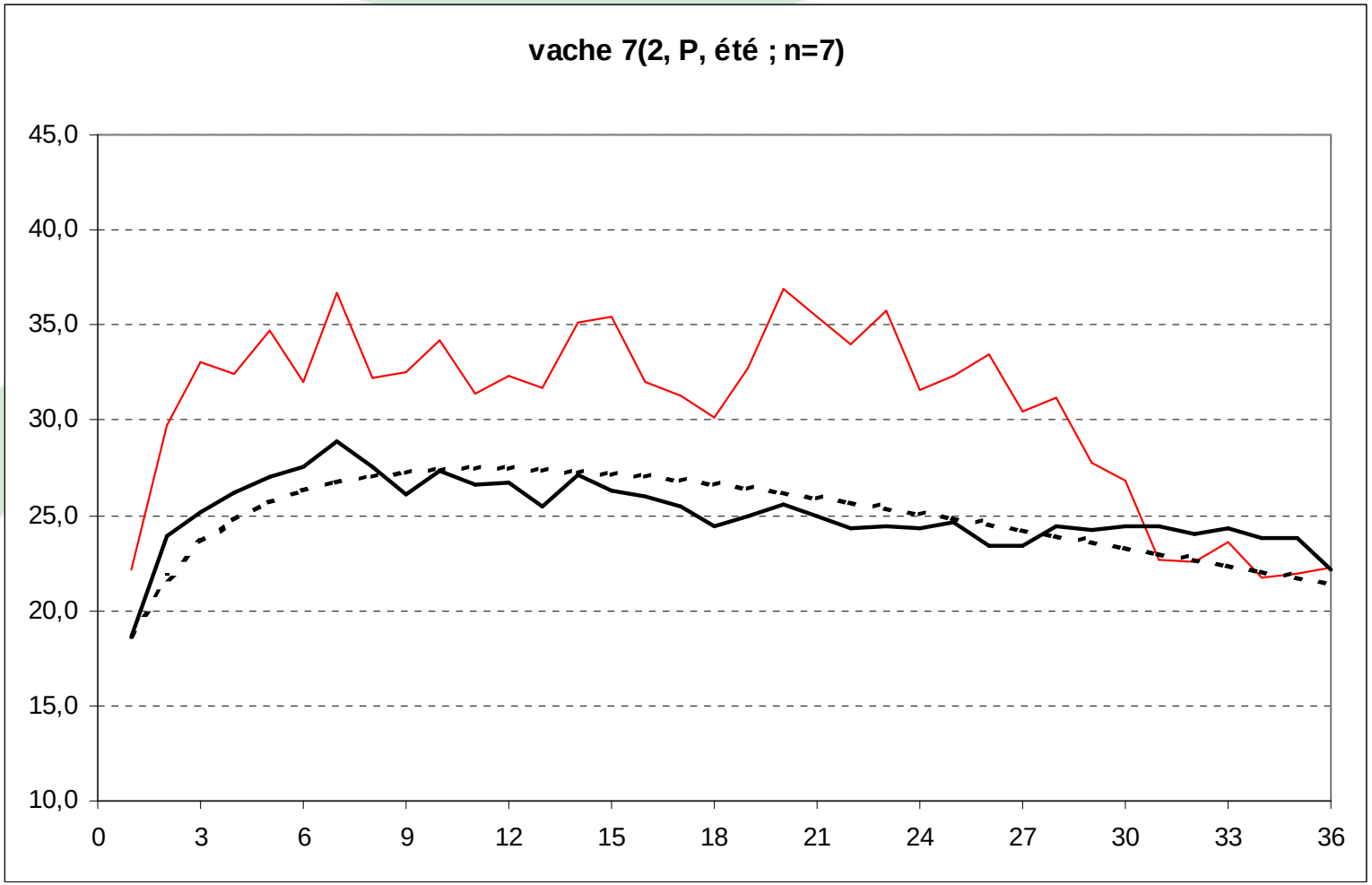
Validation du modèle


Fixed Effects Influence Diagnostics for L_LB
Restr. Likelihood Distanc
16
351
402
Deleted Level of VACHE
PRESS Statistic
367
13
316
390
Deleted Level of VACHE
Cook's D
16
351
Deleted Level of VACHE
COVRATIO
322
7
402
Deleted Level of VACHE

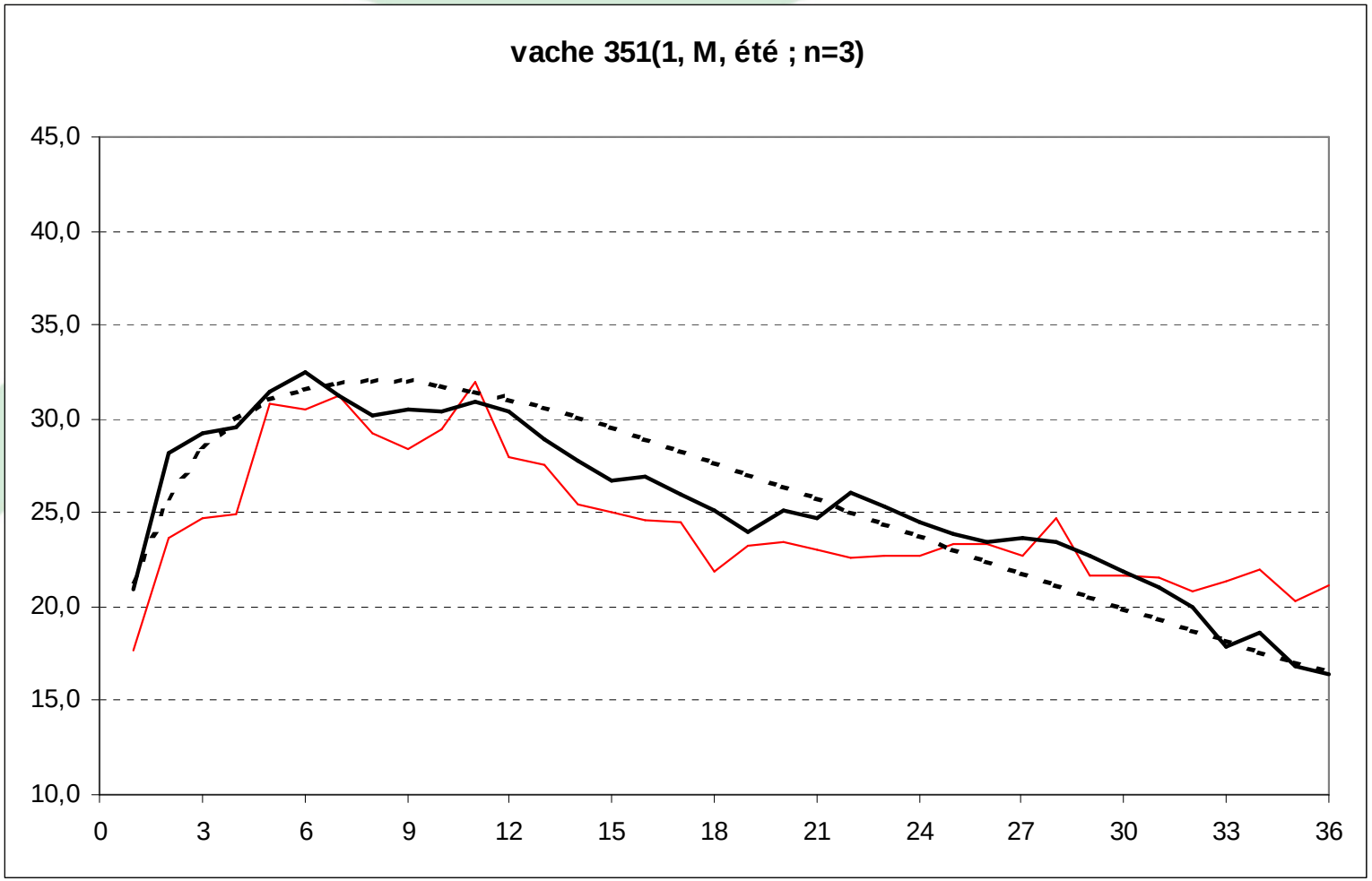
Quelques courbes « influentes »



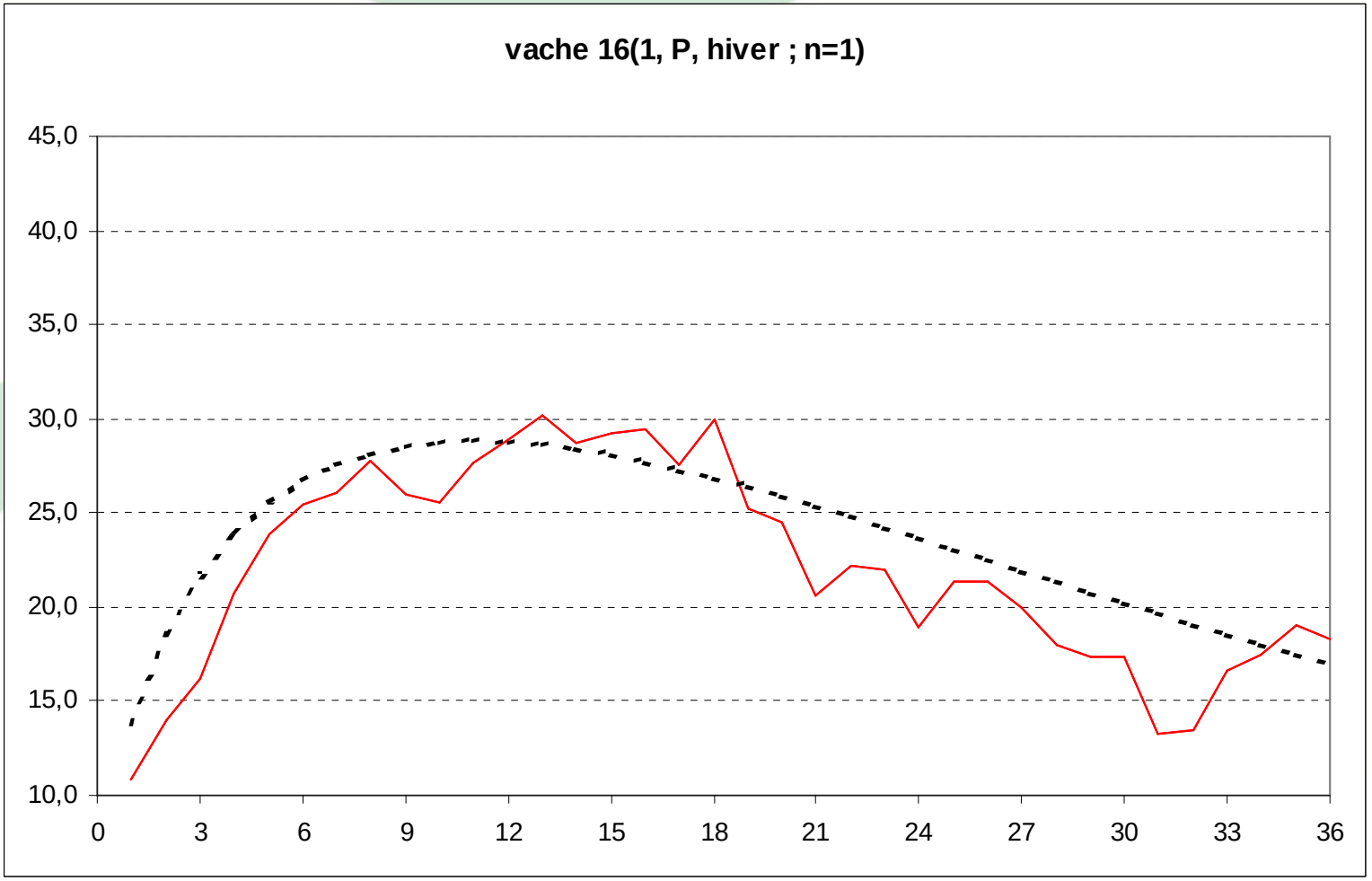
402 : Distance REML = 4,192 ; D Cook = 0,073 ; COVRATIO = 0,283 (min)



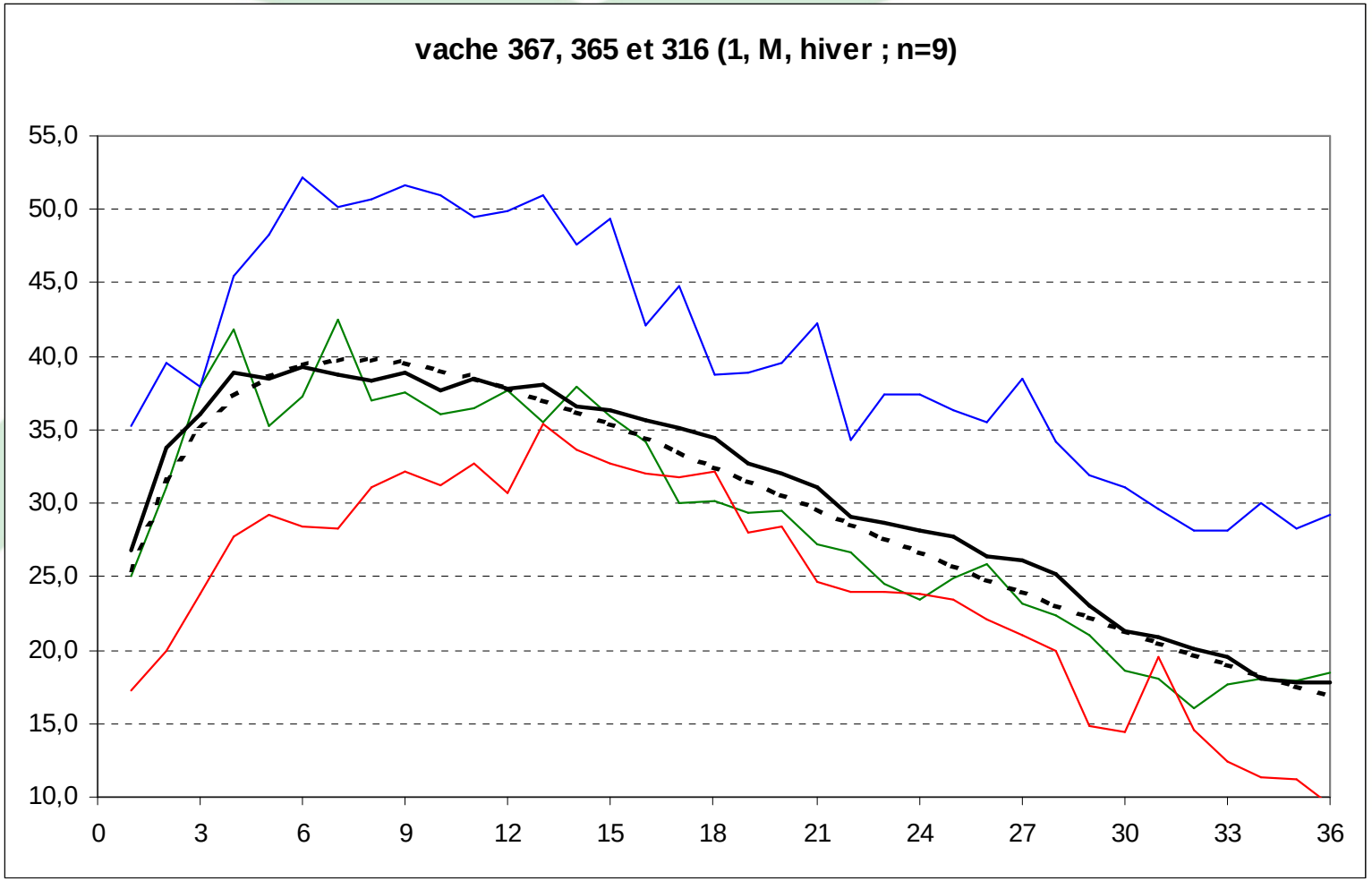
7 : Distance REML = 2,227 ;PRESS = 2,282 ; COVRATIO = 0,383



351 : Distance REML = 6,025 ; D Cook = 0,124 ; COVRATIO = 0,426



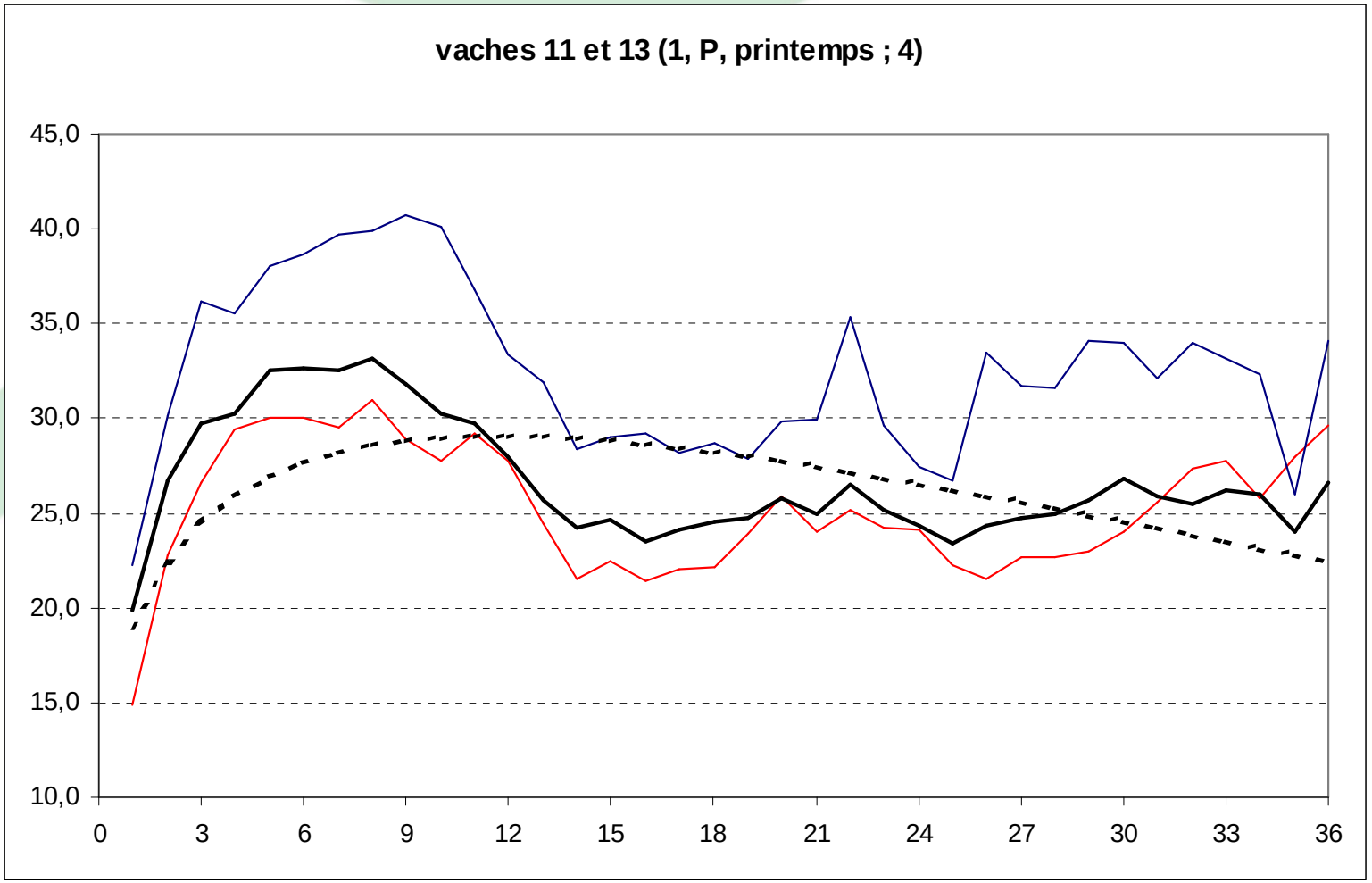
16 : Distance REML = 8,687 (max) ; PRESS = 2,147 ; D Cook = 0,248 (max)



316 : PRESS = 3,204 ; COVRATIO = 0,551

367 : PRESS = 4,685 (max) ; COVRATIO = 0,449

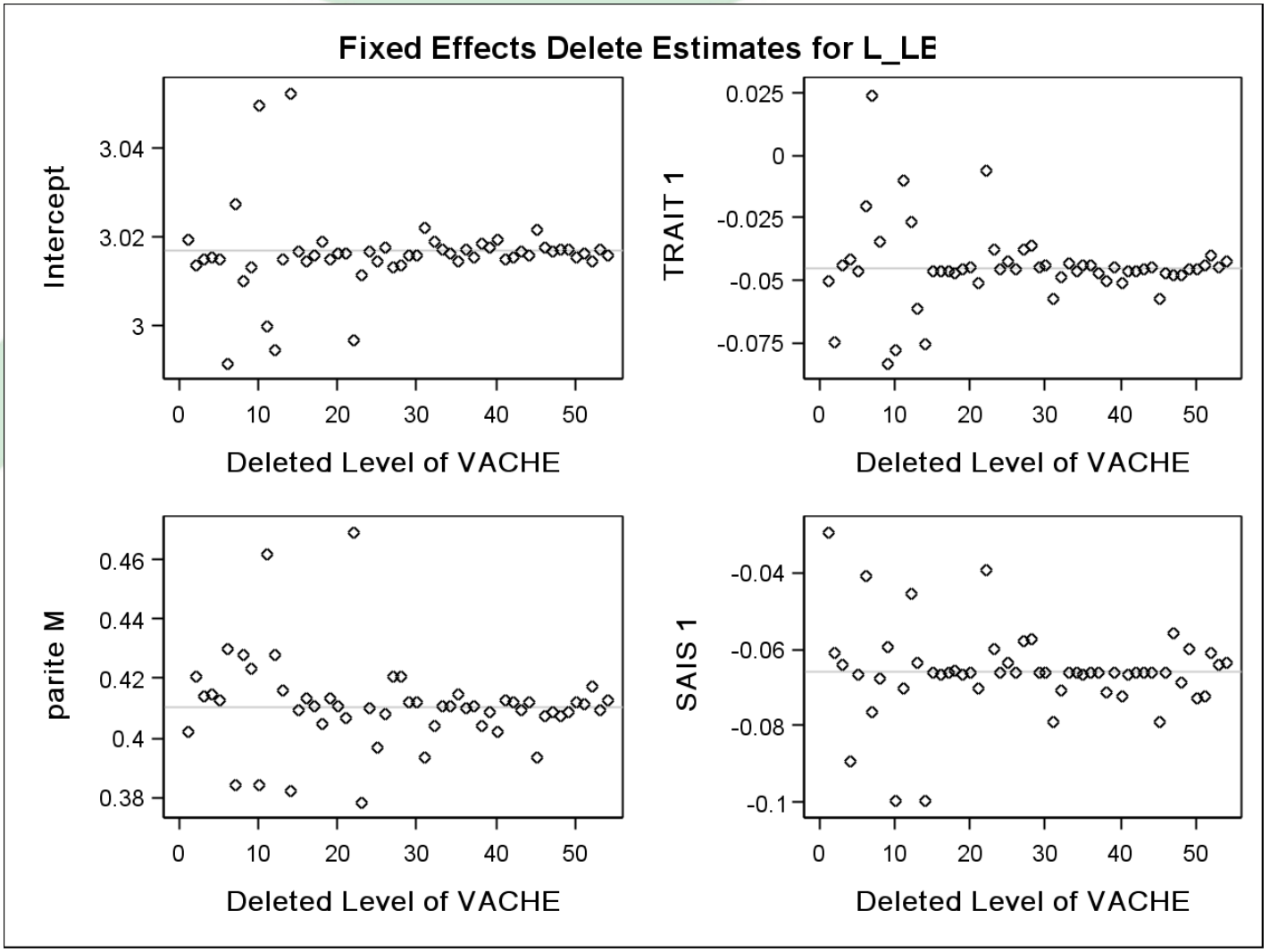
365 : Distance REML = 0,068 ; PRESS = 0,234 ; D Cook = 0,000 ; COVRATIO =2,914



13 : PRESS = 3,342

11 : Distance REML = 3,163 ; D Cook = 0,078 ; COVRATIO = 0,554

L'influence sur les paramètres des effets fixes


Fixed Effects Delete Estimates for L_LE
Intercept
3.04
3.02
3
TRAIT 1
0.025
0
-0.025
-0.05
-0.075
parite M
0.46
0.44
0.42
0.4
0.38
SAIS 1
-0.04
-0.06
-0.08
-0.1
0
10
20
30
40
50
Deleted Level of VACHE

Éléments de bibliographie

- Cook, R.D. (1986) :
Assessement of local influence, *Journal of the Royal Statistical Society*, Series B, **48**, 133-169

-Verbeke, G. et Molenberghs, G. (2000) :
Linear mixed models for Longitudinal data. New-York: Springer

- Fitzmaurice, G. M., Laird, N. M., Ware, J.H., (2004) :
Applied longitudinal analysis. John Wiley & Sons. New-York