Traitement de la multicolinéarité en régression

Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC
CNAM
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Plan

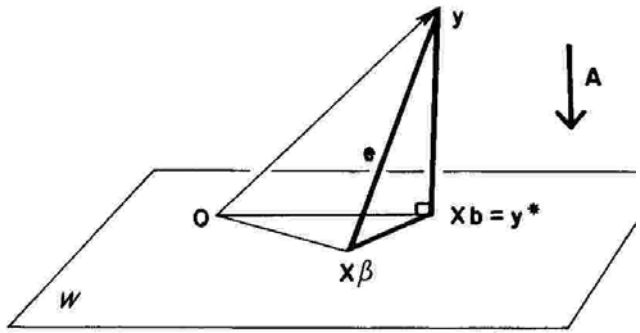
1. Rappels sur la régression multiple
2. La multicolinéarité exacte
3. Multicolinéarité approchée
4. Sélection de variables, choix de modèles
5. Régression sur composantes principales
6. Régression PLS
7. Régression ridge
8. Lasso
9. Elastic net

1. Régression linéaire multiple (rappels)

1.1 Le modèle

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

Un peu de géometrie



- Moindres carrés

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{Ay}$$

$$\mathbf{y - Xb} \perp W \quad \mathbf{(y - Xb)'Xu} = 0 \quad \forall \mathbf{u}$$

$$\mathbf{X'y = X'Xb} \quad \text{Equations normales}$$

$$\mathbf{b} = \mathbf{(X'X)^{-1}X'y}$$

$$\text{projecteur } \mathbf{A} = \mathbf{X(X'X)^{-1}X'}$$

- b estimateur de variance minimale de β parmi les estimateurs linéaires sans biais
- estimateur du maximum de vraisemblance si résidus gaussiens iid
- Estimations non uniques de β si **X'X** non inversible mais projection $\hat{\mathbf{y}}$ unique

- Variance des estimations

$$V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

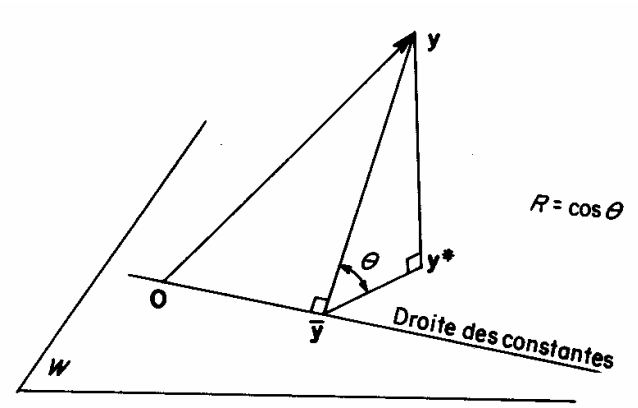
  - Estimations imprécises si multicolinéarité

- Estimation de σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1}$$

1.2 Qualité de l'ajustement

- Le R^2: cosinus carré de l'angle entre $\mathbf{y} - \bar{\mathbf{y}}$ et W



- Analyse de variance= test de nullité de R^2: absence totale de liaison

1.2 Qualité de l'ajustement (suite)

- Le R^2 est biaisé: surestimation

$$E(R^2) = \mathrm{R}^2 + \frac{p}{n-1}(1-\mathrm{R}^2) + O(\frac{1}{n^2})$$

- R^2 ajusté :

$$\hat{R}^2 = \frac{(n-1)R^2 - p}{n-p-1}$$

$$\sigma^2 = (1-R^2)s_y^2 \qquad \hat{\sigma}^2 = (1-\hat{R}^2)s^{*2}_y$$

Peut être négatif…

2. Multicolinéarité exacte

2.1 Régression sur données compositionnelles

- x_j proportions de somme =1

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon \qquad x_1 + x_2 + \ldots x_p = 1$$

$$\hat{Y} = \beta_0 (x_1 + x_2 + \ldots x_p) + \beta_1 \, x_1 + \beta_2 \, x_2 + \ldots + \beta_p \, x_p$$

$$\hat{Y} = (\beta_0 + \beta_1) x_1 + (\beta_0 + \beta_2) x_2 + \ldots . (\beta_0 + \beta_p) x_p$$

$$\hat{Y} = \alpha_1 \, x_1 + \alpha_2 \, x_2 + \ldots + \alpha_p \, x_p$$

- Régression sans constante

2.2 Régression sur variables qualitatives : le modèle linéaire général

- Un prédicteur qualitatif

Obs	NOM	PUIS	POIDS	FINITION	PRIX
1	ALFASUD-TI-1350	79	870	B	30570
2	AUDI-100-L	85	1110	TB	39990
3	SIMCA-1307-GLS	68	1050	M	29600
4	CITROEN-GS-CLUB	59	930	M	28250
5	FIAT-132-1600GLS	98	1105	B	34900
6	LANCIA-BETA-1300	82	1080	TB	35480
7	PEUGEOT-504	79	1160	B	32300
8	RENAULT-16-TL	55	1010	B	32000
9	RENAULT-30-TS	128	1320	TB	47700
10	TOYOTA-COROLLA	55	815	M	26540
11	ALFETTA-1.66	109	1060	TB	42395
12	PRINCESS-1800-HL	82	1160	B	33990
13	DATSUN-200L	115	1370	TB	43980
14	TAUNUS-2000-GL	98	1080	B	35010
15	RANCHO	80	1129	TB	39450
16	MAZDA-9295	83	1095	M	27900
17	OPEL-REKORD-L	100	1120	B	32700
18	LADA-1300	68	955	M	22100

- Recodage des modalités en indicatrices

Obs	NOM	PUIS	POIDS	F1	F2	F3	PRIX
1	ALFASUD-TI-1350	79	870	1	0	0	30570
2	AUDI-100-L	85	1110	0	0	1	39990
3	SIMCA-1307-GLS	68	1050	0	1	0	29600
4	CITROEN-GS-CLUB	59	930	0	1	0	28250
5	FIAT-132-1600GLS	98	1105	1	0	0	34900
6	LANCIA-BETA-1300	82	1080	0	0	1	35480
7	PEUGEOT-504	79	1160	1	0	0	32300
8	RENAULT-16-TL	55	1010	1	0	0	32000
9	RENAULT-30-TS	128	1320	0	0	1	47700
10	TOYOTA-COROLLA	55	815	0	1	0	26540
11	ALFETTA-1.66	109	1060	0	0	1	42395
12	PRINCESS-1800-HL	82	1160	1	0	0	33990
13	DATSUN-200L	115	1370	0	0	1	43980
14	TAUNUS-2000-GL	98	1080	1	0	0	35010
15	RANCHO	80	1129	0	0	1	39450
16	MAZDA-9295	83	1095	0	1	0	27900
17	OPEL-REKORD-L	100	1120	1	0	0	32700
18	LADA-1300	68	955	0	1	0	22100

- La somme des indicatrices vaut 1

- Estimation des paramètres indéterminée car colinéarité avec le terme constant
- Nécessité de contraintes:
  - Élimination d'une modalité (coefficient nul)

R-Square	Coeff Var	Root MSE	PRIX Mean
0.904689	6.791932	2320.030	34158.61

Parameter		Estimation	Erreur standard	t Value	Pr > |t|
Intercept		23382.59786 B	6200.788037	3.77	0.0023
PUIS		86.96368	46.069500	1.89	0.0816
POIDS		8.00795	6.568084	1.22	0.2444
FINITION	B	-6243.33612 B	1432.072306	-4.36	0.0008
FINITION	M	-10056.07842 B	1906.652796	-5.27	0.0002
FINITION	TB	0.00000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations.
Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

- Autres contraintes :
  - Coefficients à somme nulle (préférée en trade-off)
  - Solutions équivalentes car mêmes prévisions
  - Passage simple

$$\beta_1 F_1 + \beta_2 F_2 + 0F_3 = \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 0 \quad \alpha_3 = -(\alpha_1 + \alpha_2)$$

$$\alpha_1 = \beta_1 - \frac{\beta_1 + \beta_2}{3} = \frac{2}{3}\beta_1 - \frac{\beta_2}{3}$$

$$\alpha_2 = \frac{2}{3}\beta_2 - \frac{\beta_1}{3}$$

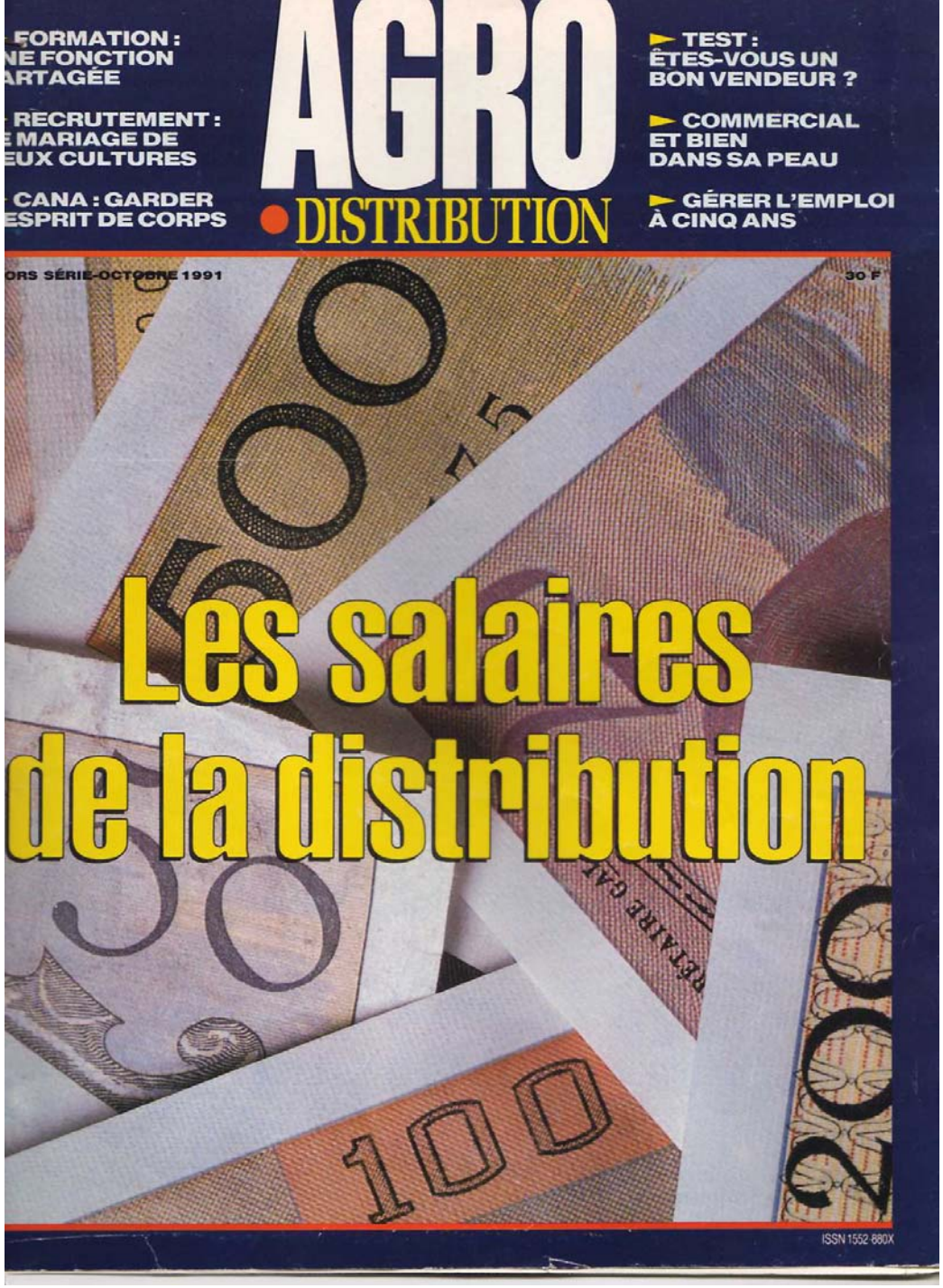
$$\alpha_3 = -\frac{\beta_1 + \beta_2}{3}$$

Les écarts ne changent pas :

$$\alpha_1 - \alpha_2 = \beta_1 - \beta_2$$

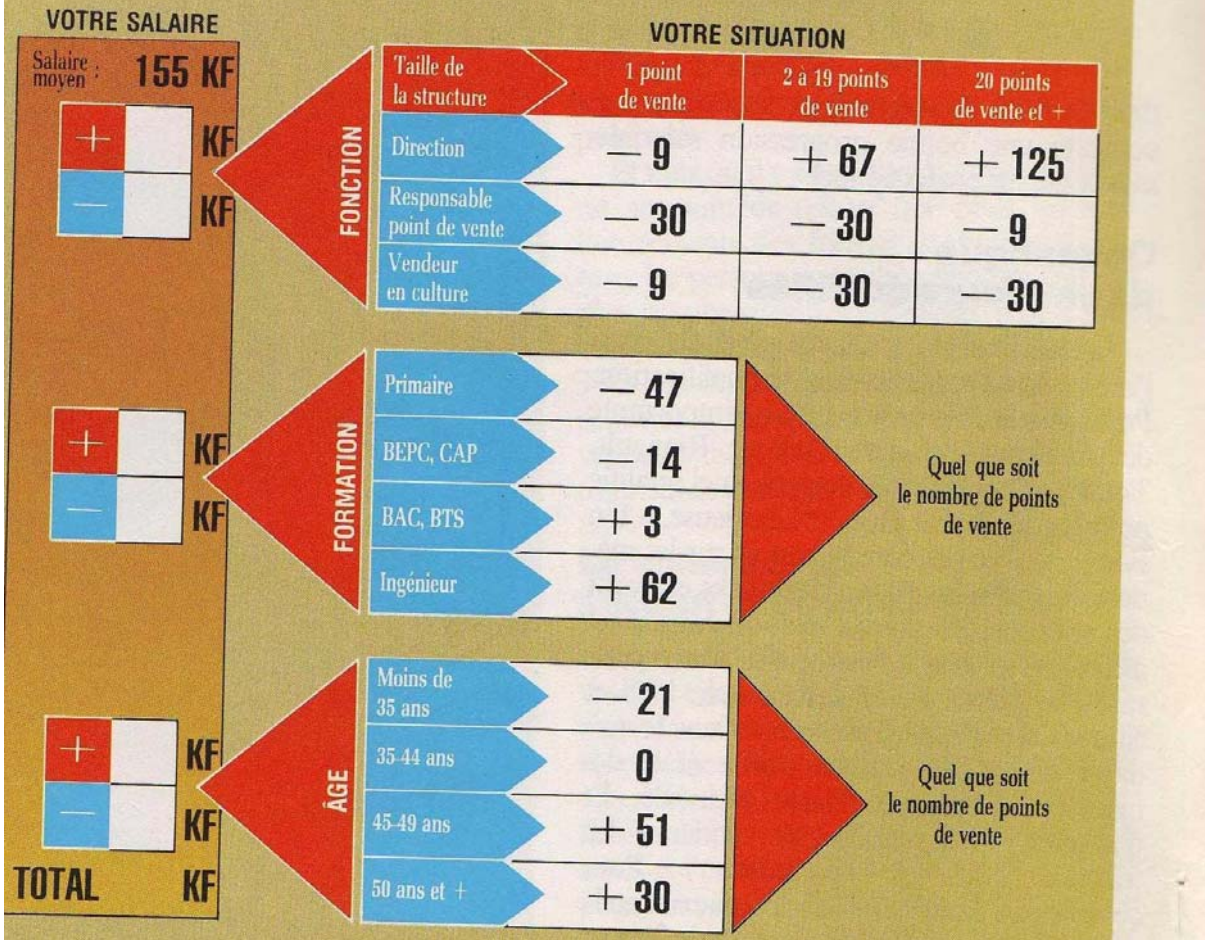
Création d'interactions

- Croisement de variables à m1 et m2 modalités:
  - Variable à m1m2 modalités


FORMATION :
NE FONCTION
ARTAGÉE
RECRUTEMENT :
E MARIAGE DE
EUX CULTURES
CANA : GARDER
ESPRIT DE CORPS
AGRO
DISTRIBUTION
TEST :
ÊTES-VOUS UN
BON VENDEUR ?
COMMERCIAL
ET BIEN
DANS SA PEAU
GÉRER L'EMPLOI
À CINQ ANS
ORS SÉRIE-OCTOBRE 1991
30 F
Les salaires
de la distribution
ISSN 1552 880X

CALCULEZ VOUS-MÊME VOTRE SALAIRE

En fonction de votre situation professionnelle, le schéma ci-dessous vous permet de calculer votre salaire. Le salaire moyen brut annuel sur l'ensemble de la profession est de 155 KF. A ce salaire, vous devez ajouter ou retrancher un certain nombre de KF selon trois critères : votre fonction, votre niveau de formation et votre âge.



Exemple de calcul : pour un directeur d'une structure comptant 8 points de vente, de formation ingénieur et âgé de 48 ans, le salaire annuel sera de $155 + (67 + 62 + 51) = 335$ KF.
L'analyse utilisée est un modèle linéaire d'analyse de variance. La corrélation entre les salaires réels et ceux trouvés par la formule est de 0,69.

```
data;
length age $ 12 ft $ 12  diplome $ 12 fonction $ 12 ;
title ' modele avec interaction fonction taille complete';
infile 'c:\GILBERT\BVA\agrod.dat';
input numero $ 10-13 fonct  16 type $ 25 pvente 26-28
      salb 36-38 salcod 39 age1 74  dipl1 75;
if pvente = 1 then taille = 1;
if 2<=pvente <=19 then taille = 2;
if pvente >=20 then taille = 3;
sal = 25 + 50*(salcod - 1);
if sal >= 25 then salaire = sal;
else salaire = salb;
if age1<=2 then   age='<34';
if 3<= age1 <=4 then age='35-44';
if age1=5 then age='45-49';
if age1>=6 then age='>50' ;
if dipl1=1 then diplome='Primaire';
if dipl1=2 then diplome='Bepc';
if dipl1=3 or 5<= dipl1<=6 then diplome='Bac ou BTS';
if dipl1=4 then diplome='CAP';
if dipl1=7 then diplome='Ingenieur';
if fonct=1 then fonction ='Directeur';
if 2<=fonct<=3 then fonction='Autre';
if fonct>=2 then ft= 'autre';
if fonct=1 and taille=1 then ft='dir1';
if fonct=1 and taille=2 then ft ='dir2';
if fonct=1 and taille=3 then ft ='dir3';

proc glm;
class fonct taille age  diplome;
model salaire = fonct*taille age diplome /solution  p cli;
lsmeans fonct*taille age diplome /p;
run;
```

The GLM Procedure

Dependent Variable: salaire

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	15	1100307.201	73353.813	15.71	<.0001
Error	262	1222952.616	4667.758		
Corrected Total	277	2323259.817			

R-Square	Coeff Var	Root MSE	salaire Mean
0.473605	44.83449	68.32099	152.3849

Source	DDL	Type I SS	Carré moyen	Valeur F	Pr > F
fonct*taille	8	871474.6291	108934.3286	23.34	<.0001
age	3	98959.5435	32986.5145	7.07	0.0001
diplome	4	129873.0280	32468.2570	6.96	<.0001

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
fonct*taille	8	325123.6376	40640.4547	8.71	<.0001
age	3	139366.8734	46455.6245	9.95	<.0001
diplome	4	129873.0280	32468.2570	6.96	<.0001

Parameter		Estimation	standard	t Value	Pr > |t|
Intercept		115.8767770 B	18.47798316	6.27	<.0001
fonct*taille 1 1		19.2621518 B	17.11380134	1.13	0.2614
fonct*taille 1 2		92.8556242 B	18.00924814	5.16	<.0001
fonct*taille 1 3		150.7352311 B	27.60425699	5.46	<.0001
fonct*taille 2 1		-19.8378255 B	22.17433682	-0.89	0.3718
fonct*taille 2 2		-8.5372771 B	16.10439175	-0.53	0.5965
fonct*taille 2 3		17.4229154 B	16.81984076	1.04	0.3012
fonct*taille 3 1		11.4642504 B	20.12480471	0.57	0.5694
fonct*taille 3 2		0.3273331 B	16.08346634	0.02	0.9838
fonct*taille 3 3		0.0000000 B	.	.	.
age	35-44	-31.4722782 B	14.19292624	-2.22	0.0274
age	45-49	20.2791785 B	16.66074286	1.22	0.2246
age	<34	-51.7826584 B	14.98593828	-3.46	0.0006
age	>50	0.0000000 B	.	.	.
diplome	Bac ou BTS	46.5772677 B	16.01503767	2.91	0.0039
diplome	Bepc	24.4751634 B	17.70833447	1.38	0.1681
diplome	CAP	39.0939164 B	19.57272489	2.00	0.0468
diplome	Ingenieur	107.2743816 B	20.86902338	5.14	<.0001
diplome	Primaire	0.0000000 B	.	.	.

The GLM Procedure

Least Squares Means

fonct	taille	salaire LSMEAN	LSMEAN Number
1	1	162.879135	1
1	2	236.472608	2
1	3	294.352214	3
2	1	123.779158	4
2	2	135.079706	5
2	3	161.039899	6
3	1	155.081234	7
3	2	143.944316	8
3	3	143.616983	9

Least Squares Means for effect fonct*taille
Pr > |t| for HO: LSMean(i)=LSMean(j)

Dependent Variable: salaire

i/j	1	2	3	4	5	6	7	8	9
1		<.0001	<.0001	0.0659	0.0720	0.9078	0.6947	0.2321	0.2614
2	<.0001		0.0238	<.0001	<.0001	<.0001	0.0001	<.0001	<.0001
3	<.0001	0.0238		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
4	0.0659	<.0001	<.0001		0.5883	0.0804	0.1937	0.3428	0.3718
5	0.0720	<.0001	<.0001	0.5883		0.0940	0.2935	0.5477	0.5965
6	0.9078	<.0001	<.0001	0.0804	0.0940		0.7604	0.2712	0.3012
7	0.6947	0.0001	<.0001	0.1937	0.2935	0.7604		0.5559	0.5694
8	0.2321	<.0001	<.0001	0.3428	0.5477	0.2712	0.5559		0.9838
9	0.2614	<.0001	<.0001	0.3718	0.5965	0.3012	0.5694	0.9838	

age	LSMEAN	Number
35-44	157.187801	1
45-49	208.939257	2
<34	136.877420	3
>50	188.660079	4

Least Squares Means for effect age
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: salaire

i/j	1	2	3	4
1		0.0003	0.0540	0.0274
2	0.0003		<.0001	0.2246
3	0.0540	<.0001		0.0006
4	0.0274	0.2246	0.0006	

diplome	LSMEAN	Number
Bac ou BTS	176.009261	1
Bepc	153.907157	2
CAP	168.525910	3
Ingenieur	236.706375	4
Primaire	129.431993	5

Least Squares Means for effect diplome
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: salaire

i/j	1	2	3	4	5
1		0.1174	0.6103	0.0002	0.0039
2	0.1174		0.4286	<.0001	0.1681
3	0.6103	0.4286		0.0011	0.0468
4	0.0002	<.0001	0.0011		<.0001
5	0.0039	0.1681	0.0468	<.0001	

Deuxieme modèle

```
if fonct=3 and taille=1 then ft='1';
if fonct=2 and taille=3 then ft='1';
if fonct=1 and taille=1 then ft='1';
if fonct=1 and taille=2 then ft ='2';
if fonct=1 and taille=3 then ft ='3';
if fonct=2 and taille=1 then ft='4';
if fonct=2 and taille=2 then ft='4';
if fonct=3 and taille=2 then ft='4';
if fonct=3 and taille=3 then ft='4';

proc glm;
class ft  age  diplome;
model salaire = ft age diplome /solution  p clm;
lsmeans ft age diplome /p;
run;
```

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	9	1091494.221	121277.136	26.39	<.0001
Error	268	1231765.596	4596.140		
Corrected Total	277	2323259.817			

R-Square	Coeff Var	Root MSE	salaire Mean
0.469812	44.48921	67.79484	152.3849

Parameter		Estimation		Erreur standard	t Value	Pr > |t|
Intercept		108.1801431	B	13.79447986	7.84	<.0001
ft	1	20.9719018	B	9.20765648	2.28	0.0235
ft	2	96.7800140	B	14.04629291	6.89	<.0001
ft	3	155.3020627	B	24.88831243	6.24	<.0001
ft	4	0.0000000	B	.	.	.
age	35-44	-29.9167734	B	13.87680050	-2.16	0.0320
age	45-49	21.3891613	B	16.35408144	1.31	0.1920
age	<34	-50.8856349	B	14.40849511	-3.53	0.0005
age	>50	0.0000000	B	.	.	.
diplome	Bac ou BTS	49.5558603	B	15.32114064	3.23	0.0014
diplome	Bepc ou CAP	32.8345809	B	15.77742828	2.08	0.0384
diplome	Ingenieur	109.2620741	B	20.42227391	5.35	<.0001
diplome	Primaire	0.0000000	B	.	.	.

The GLM Procedure

Observation		Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value	
1		75.0000000	127.8222703	-52.8222703	109.1875210	146.4570195
2		125.0000000	129.1520449	-4.1520449	102.1480528	156.1560369
3		75.0000000	129.1520449	-54.1520449	102.1480528	156.1560369
4		75.0000000	148.7911318	-73.7911318	128.3923441	169.1899195
5		125.0000000	254.5160174	-129.5160174	221.9562145	287.0758203
6		120.0000000	127.8222703	-7.8222703	109.1875210	146.4570195
7		120.0000000	262.1524312	-142.1524312	211.9482235	312.3566390
8		75.0000000	200.0970665	-125.0970665	172.4353924	227.7587406
9		225.0000000	237.7947380	-12.7947380	202.0904884	273.4989876
10		225.0000000	150.5412062	74.4587938	116.1014683	184.9809442
11		175.0000000	175.0433837	-0.0433837	135.8599746	214.2267928
12		350.0000000	314.2222312	35.7777688	275.7462426	352.6982198
13		125.0000000	148.7911318	-23.7911318	128.3923441	169.1899195
14		216.0000000	204.9601571	11.0398429	170.0434036	239.8769106
15		75.0000000	148.7911318	-73.7911318	128.3923441	169.1899195
16		225.0000000	208.4973456	16.5026544	174.8432211	242.1514700
17		125.0000000	148.7911318	-23.7911318	128.3923441	169.1899195
18		125.0000000	129.1520449	-4.1520449	102.1480528	156.1560369
19	*	.	161.9866258	.	133.4713076	190.5019440
20		170.0000000	224.5992440	-54.5992440	198.3268148	250.8716732
21		400.0000000	372.7442799	27.2557201	325.2811190	420.2074408
22		75.0000000	111.1009909	-36.1009909	88.0153232	134.1866586
23		150.0000000	284.3054578	-134.3054578	250.9978248	317.6130907
24		325.0000000	275.9051787	49.0948213	243.6405200	308.1698375
25		120.0000000	148.7911318	-28.7911318	128.3923441	169.1899195

3 La multicolinéarité approchée

3.1 Un exemple

Obs	NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE	NAT	FINITION	PRIX
1	ALFASUD-TI-1350	1350	79	393	161	870	165	I	B	30570
2	AUDI-100-L	1588	85	468	177	1110	160	D	TB	39990
3	SIMCA-1307-GLS	1294	68	424	168	1050	152	F	M	29600
4	CITROEN-GS-CLUB	1222	59	412	161	930	151	F	M	28250
5	FIAT-132-1600GLS	1585	98	439	164	1105	165	I	B	34900
6	LANCIA-BETA-1300	1297	82	429	169	1080	160	I	TB	35480
7	PEUGEOT-504	1796	79	449	169	1160	154	F	B	32300
8	RENAULT-16-TL	1565	55	424	163	1010	140	F	B	32000
9	RENAULT-30-TS	2664	128	452	173	1320	180	F	TB	47700
10	TOYOTA-COROLLA	1166	55	399	157	815	140	J	M	26540
11	ALFETTA-1.66	1570	109	428	162	1060	175	I	TB	42395
12	PRINCESS-1800-HL	1798	82	445	172	1160	158	GB	B	33990
13	DATSUN-200L	1998	115	469	169	1370	160	J	TB	43980
14	TAUNUS-2000-GL	1993	98	438	170	1080	167	D	B	35010
15	RANCHO	1442	80	431	166	1129	144	F	TB	39450
16	MAZDA-9295	1769	83	440	165	1095	165	J	M	27900
17	OPEL-REKORD-L	1979	100	459	173	1120	173	D	B	32700
18	LADA-1300	1294	68	404	161	955	140	U	M	22100

- *Résultats*

Analysis of Variance

Source	DDL	Sum of Squares	Mean Square	Valeur F	Pr > F
Model	6	520591932	86765322	4.47	0.0156
Error	11	213563858	19414896		
Corrected Total	17	734155790			

Root MSE	4406.23379	R-Square	0.7091
Dependent Mean	34159	Adj R-Sq	0.5504
Coeff Var	12.89934		

Parameter Estimates

Variable	DDL	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-8239.36268	42718	-0.19	0.8506	0
CYL	1	-3.50518	5.55060	-0.63	0.5406	3.77201
PUIS	1	282.16880	174.88297	1.61	0.1349	11.11882
LON	1	-15.03766	129.74749	-0.12	0.9098	7.20420
LAR	1	208.69377	412.04788	0.51	0.6225	4.19760
POIDS	1	12.57468	24.62219	0.51	0.6197	9.95728
VITESSE	1	-111.11355	222.25657	-0.50	0.6270	6.37511

3.2 estimation et test des coefficients

- Dans l'exemple aucun coefficient significatif!
- Le test de Student d'un coefficient n'est pas un test de non corrélation, mais un test de non apport d'une variable conditionnellement aux p-1 autres
- Explication: la multicolinéarité
- De fortes corrélations entre prédicteurs conduisent à de mauvaises estimations des b_j car $\det(X'X) \approx 0$

3.3 Détection

- Etude de la matrice de corrélation

Matrice de corrélations

	CYL	PUIS	LON	LAR	POIDS	VITESSE
CYL	1.00000	0.79663	0.70146	0.62976	0.78895	0.66493
PUIS	0.79663	1.00000	0.64136	0.52083	0.76529	0.84438
LON	0.70146	0.64136	1.00000	0.84927	0.86809	0.47593
LAR	0.62976	0.52083	0.84927	1.00000	0.71687	0.47295
POIDS	0.78895	0.76529	0.86809	0.71687	1.00000	0.47760
VITESSE	0.66493	0.84438	0.47593	0.47295	0.47760	1.00000

- ACP (voir plus loin)
- Analyse des facteurs d'inflation de la variance

$$V(b_j) = \frac{\sigma^2}{n}\mathbf{R}^{-1}_{j,j} = \frac{\sigma^2}{n}\frac{1}{1-R^2(x_j; x_1, x_2, ..., x_p)} = \frac{\sigma^2}{n}VIF$$

4. Sélection de variables, choix de modèles

- Choix de k variables parmi p
  - Elimination de variables non pertinentes
  - Obtention de formules plus stables
- Critères:
  - A k fixé: R^2
  - Comparaison de modèles de tailles différentes
    - Capacité prédictive, complexité

la recherche de la parsimonie: le rasoir d'Ockham

Guillaume d'Occam (1285? – 1349?), dit le « docteur invincible » franciscain philosophe logicien et théologien scolastique.
Etudes à Oxford, puis Paris. Enseigne quelques années à Oxford. Accusé d'hérésie, convoqué pour s'expliquer à Avignon, excommunié pour avoir fui à Munich à la cour de Louis IV de Bavière. Meurt vraisemblablement de l'épidémie de peste noire.

Principe de raisonnement attribué à Occam : « Les multiples ne doivent pas être utilisés sans nécessité » (*pluralitas non est ponenda sine necessitate*).

A inspiré le personnage du moine franciscain Guillaume de Baskerville dans le « Nom de la rose » d'Umberto Eco. *Premier jour, vêpres : « il ne faut pas multiplier les explications et les causes sans qu'on en ait une stricte nécessité. »*

4.1 Quelques critères pour comparer des modèles de tailles différentes

- R^2 ajusté et $\hat{\sigma}$ sont équivalents

- AIC (Akaiké) et BIC (Schwartz)

$$AIC = -2\ln(L) + 2(k+1) = n\ln(\frac{SSE}{n}) + 2(k+1) + n(\ln\pi + 1)$$

$$BIC = -2\ln(L) + \ln(n)(k+1) = n\ln(\frac{SSE}{n}) + \ln(n)(k+1) + n(\ln\pi + 1)$$

  - On cherche à minimiser AIC ou BIC

- *SSE*= somme des carrés des résidus du modèle à k variables

Le C_p de Mallows

- On cherche à estimer l'erreur quadratique de prédiction (MSPE) :

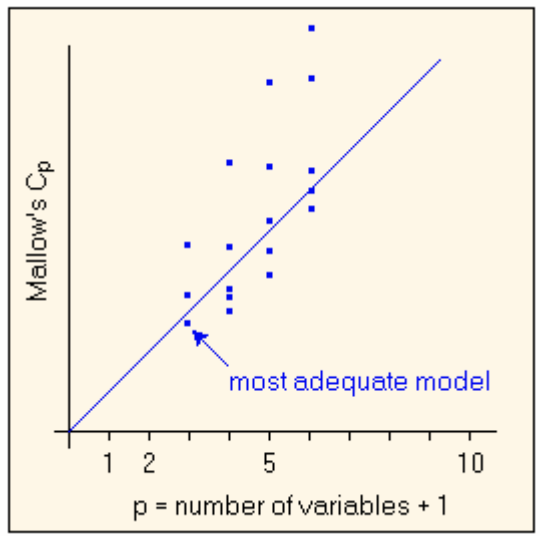
$$E\sum_j(\hat{Y}_j - E(Y_j|X_j))^2/\sigma^2,$$

- En sélectionnant P prédicteurs parmi K :

$$C_p = \frac{SSE_p}{S^2} - N + 2P,$$

- Si le modèle est le bon $E(C_p)=p$

The general procedure to find an adequate model by means of the Cp statistic is to calculate Cp for all possible combinations of variables and the Cp values against p. The model with the lowest Cp value approximately equal to p is the most "adequate" model.



http://www.statistics4u.com/fundstat_eng/cc_varsel_mallowscp.html

4.2 algorithmes de sélection

- Dénombrement exhaustif: 2^p-1 modèles
- Meilleurs sous-ensembles (algorithme de Furnival et Wilson jusqu'à quelques dizaines de variables)
- Méthodes pas à pas (stepwise selection)
  - Ascendant (forward)
  - Descendant (backward)
  - Ascendant avec élimination possible (stepwise)
  - …

Les logiciels classiques utilisent des tests d'arret :
F pour entrer, pour rester

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	Valeur F	Pr > F
1	PUIS	1	0.6379	0.6379	-0.3084	28.19	<.0001
2	POIDS	2	0.0487	0.6866	-0.1501	2.33	0.1476

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	Valeur F	Pr > F
1	LON	5	0.0004	0.7087	5.0134	0.01	0.9098
2	VITESSE	4	0.0069	0.7018	3.2760	0.29	0.6025
3	LAR	3	0.0030	0.6988	1.3900	0.13	0.7228
4	CYL	2	0.0122	0.6866	-0.1501	0.57	0.4646

- Les méthodes basées sur des tests F devraient être abandonnées:
  - utilisation incorrecte de tests multiples
  - erreurs standard ne tenant pas compte du processus de sélection
- R^2 augmente avec le nombre de variables mais pas R^2 ajusté

```
proc reg;
title Regression OLS;
id nom;
model prix=cyl puis lon lar poids vitesse/AIC BIC ADJRSQ selection=ADJRSQUARE ;
run;
```

Adjusted R-Square Selection Method

Nombre dans le modèle	R carré ajusté	R-carré	AIC	BIC	Variables du modèle
2	0.6366	0.6820	285.0800	289.8602	PUIS POIDS
3	0.6241	0.6946	286.3954	292.6816	CYL PUIS POIDS
3	0.6196	0.6909	286.5979	292.7777	PUIS LAR VITESSE
3	0.6169	0.6888	286.7168	292.8340	PUIS POIDS VITESSE
3	0.6156	0.6877	286.7743	292.8612	PUIS LON VITESSE
1	0.6137	0.6379	285.2916	288.5249	PUIS
2	0.6136	0.6619	286.1222	290.5380	PUIS VITESSE
2	0.6128	0.6612	286.1601	290.5624	PUIS LON
3	0.6098	0.6829	287.0320	292.9829	PUIS LAR POIDS
3	0.6089	0.6823	287.0683	293.0000	PUIS LON POIDS
2	0.6077	0.6567	286.3822	290.7054	PUIS LAR
4	0.5968	0.6976	288.2281	295.9091	CYL PUIS POIDS VITESSE
4	0.5967	0.6975	288.2312	295.9099	CYL PUIS LAR POIDS
4	0.5952	0.6964	288.2918	295.9270	CYL PUIS LAR VITESSE
4	0.5940	0.6955	288.3449	295.9420	PUIS LAR POIDS VITESSE
4	0.5938	0.6953	288.3533	295.9443	CYL PUIS LON POIDS
4	0.5902	0.6926	288.5046	295.9871	PUIS LON LAR VITESSE
3	0.5897	0.6666	287.8850	293.3848	CYL PUIS LON
4	0.5894	0.6920	288.5366	295.9962	CYL PUIS LON VITESSE
4	0.5894	0.6920	288.5377	295.9965	PUIS LON POIDS VITESSE
2	0.5864	0.6381	287.2815	291.2802	CYL PUIS
3	0.5839	0.6620	288.1210	293.4958	CYL PUIS VITESSE
3	0.5839	0.6619	288.1216	293.4961	CYL PUIS LAR
3	0.5839	0.6619	288.1230	293.4968	PUIS LON LAR
4	0.5773	0.6830	289.0285	296.1362	PUIS LON LAR POIDS
2	0.5761	0.6291	287.6967	291.5440	POIDS VITESSE
5	0.5721	0.7058	289.7570	299.1768	CYL PUIS LAR POIDS VITESSE
5	0.5648	0.7008	290.0464	299.2005	CYL PUIS LON POIDS VITESSE
5	0.5619	0.6988	290.1594	299.2101	CYL PUIS LON LAR VITESSE
5	0.5603	0.6977	290.2213	299.2155	CYL PUIS LON LAR POIDS

Nombre dans le modèle	R carré ajusté	R-carré	AIC	BIC	Variables du modèle
1	0.5586	0.5862	287.5579	290.3050	POIDS
4	0.5577	0.6683	289.8011	296.3600	CYL PUIS LON LAR
5	0.5572	0.6956	290.3398	299.2259	PUIS LON LAR POIDS VITESSE
3	0.5502	0.6346	289.4448	294.1209	CYL POIDS VITESSE
3	0.5455	0.6307	289.6222	294.2052	LAR POIDS VITESSE
3	0.5450	0.6303	289.6429	294.2151	LON POIDS VITESSE
6	0.5293	0.7058	291.7567	302.5767	CYL PUIS LON LAR POIDS VITESSE
2	0.5290	0.5879	289.4883	292.6779	CYL POIDS
2	0.5278	0.5869	289.5308	292.7048	LON POIDS
2	0.5272	0.5863	289.5553	292.7203	LAR POIDS
4	0.5146	0.6360	291.3804	296.8377	CYL LAR POIDS VITESSE
4	0.5146	0.6359	291.3821	296.8382	CYL LON POIDS VITESSE
4	0.5078	0.6308	291.6176	296.9123	LON LAR POIDS VITESSE
3	0.4933	0.5883	291.4700	295.0959	CYL LON POIDS
3	0.4928	0.5879	291.4882	295.1048	CYL LAR POIDS
3	0.4917	0.5870	291.5235	295.1221	LON LAR POIDS
5	0.4709	0.6362	293.3685	299.5832	CYL LON LAR POIDS VITESSE
4	0.4515	0.5886	293.4578	297.5212	CYL LON LAR POIDS
2	0.4341	0.5048	292.6099	294.6691	LON VITESSE
2	0.4166	0.4896	293.1266	295.0038	CYL LON
3	0.4074	0.5185	294.1329	296.4390	CYL LON VITESSE
1	0.3979	0.4355	292.8372	294.3548	LON
3	0.3947	0.5082	294.4956	296.6288	LON LAR VITESSE
3	0.3758	0.4928	295.0167	296.9045	CYL LON LAR
2	0.3732	0.4516	294.3463	295.8018	CYL VITESSE
4	0.3629	0.5222	296.0034	298.4649	CYL LON LAR VITESSE
2	0.3576	0.4379	294.7664	296.0795	LON LAR
1	0.3566	0.3969	293.9634	295.2214	CYL
2	0.3543	0.4350	294.8529	296.1369	LAR VITESSE
3	0.3523	0.4737	295.6452	297.2421	CYL LAR VITESSE
2	0.3511	0.4322	294.9357	296.1919	CYL LAR
1	0.3330	0.3747	294.5769	295.6955	VITESSE
1	0.2380	0.2856	296.8401	297.4603	LAR

4.3 Sur les critères de choix de modèles

- AIC et BIC ne sont semblables qu'en apparence
- **Théories différentes**
  - AIC : approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée

$$I(f;g) = \int f(t) \ln \frac{f(t)}{g(t)} dt = E_f(\ln(f(t)) - E_f(\ln(g(t))$$

Asymptotiquement:

$$E_{\hat{\theta}} E_f(\ln(g(t;\hat{\theta})) \sim \ln(L(\hat{\theta})) - k$$

BIC : choix bayesien de modèles

m modèles M_i paramétrés par θ_i de probabilités *a priori* $P(M_i)$ égales. Distribution *a priori* de θ_i pour chaque modèle $P(\theta_i / M_i)$.

Distribution *a posteriori* du modèle sachant les données $P(x/M_i)$ ou vraisemblance intégrée

Choix du modèle le plus probable *a posteriori* revient à maximiser

$$\ln(P(\mathbf{x}/M_i) \sim \ln(P(\mathbf{x}/\hat{\theta}_i, M_i) - \frac{k}{2}\ln(n)$$

$$P(M_i/\mathbf{x}) = \frac{e^{-0.5BIC_i}}{\sum_{j=1}^{m} e^{-0.5BIC_j}}$$

Comparaison AIC BIC

- Si *n* tend vers l'infini la probabilité que le *BIC* choisisse le vrai modèle tend vers 1, ce qui est faux pour l'*AIC*.
- *AIC* va choisir le modèle qui maximisera la vraisemblance de futures données et réalisera le meilleur compromis biais-variance
- L'*AIC* est un critère prédictif tandis que le *BIC* est un critère explicatif.
- Pour *n* fini: résultats contradictoires. *BIC* ne choisit pas toujours le vrai modèle: il a tendance à choisir des modèles trop simples en raison de sa plus forte pénalisation
- **Illogisme à utiliser les deux simultanément**

4.4 Ajuster ou prédire?

- Les critères précédents utilisent deux fois les données: une fois pour estimer, une autre pour mesurer la qualité
- Prédire les données futures et non le passé!
- Minimiser l'espérance de l'erreur quadratique de prédiction $E(y-\hat{y})^2$

- Solution pratique: la validation croisée
  - « Leave one out »: chaque observation est estimée à l'aide des n-1 autres
    - résidu prédit:

$$y_i - \hat{f}^{(-i)}(x_i) = y_i - \hat{y}_i^{(-i)} = \frac{y_i - \hat{y}_i}{1 - h_i}$$

h_i terme diagonal du projecteur $\mathbf{X(X'X)^{-1}X'}$

    - PRESS predicted error sum of squares

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(-i)} \right)^2$$

quelques press

- modéle complet: 732726946
- puissance poids 308496438
- cylindree puissance poids 369112558
- puissance 327142373

4.5 Sélectionner ou non?

- Contestable si on a un modèle: difficile de proposer à l'utilisateur une formule qui ne tient pas compte de variables pourtant influentes et ne permet pas de quantifier l'effet de leurs variations sur la réponse Y.

Coefficients de corrélation de Pearson, N = 18

	CYL	PUIS	LON	LAR	POIDS	VITESSE
PRIX	0.63858	0.79870	0.64376	0.54665	0.75329	0.58176

Comment garder toutes les variables?

- Régression sur composantes principales
- Régression PLS
- Régression ridge
- Lasso

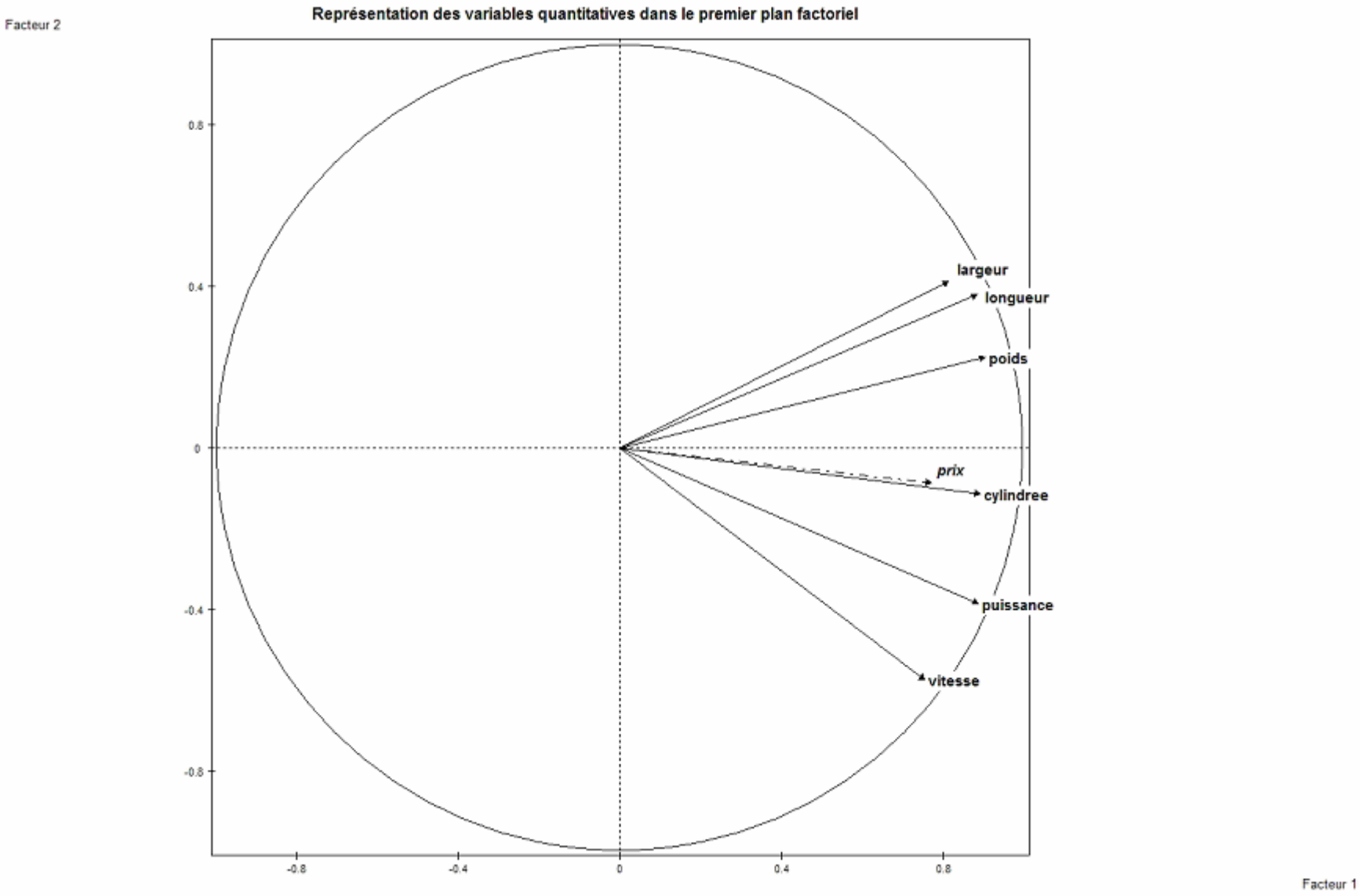
- Utile pour le cas maudit: p>n
- Mais: perte de certaines propriétés: estimateurs biaisés, non-invariance par changement d'échelle
- Nécessité de centrer réduire au préalable

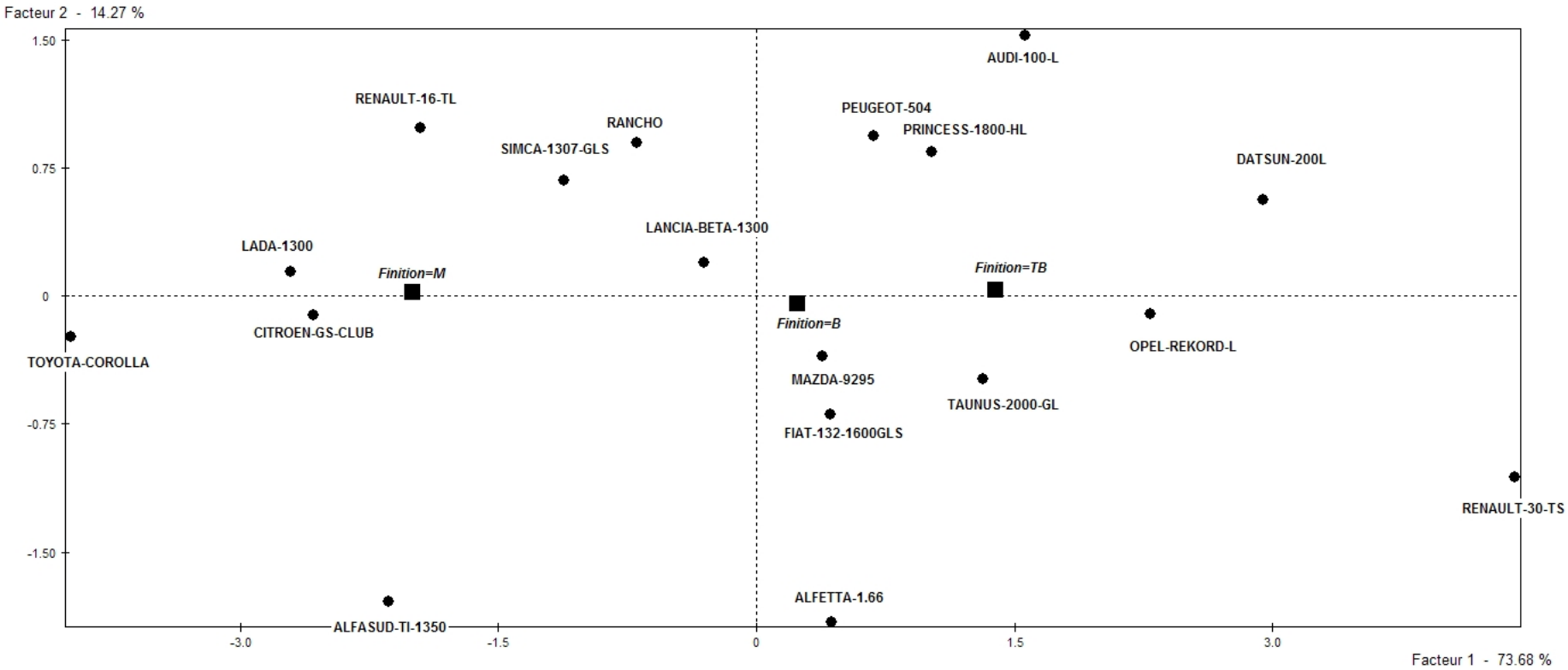
5. Régression sur composantes principales

- ACP sur les X
- Chaque composante est une combinaison linéaire de tous les prédicteurs
- Régression ascendante sur la première composante, puis sur les deux premieres etc.
- Composantes principales non corrélées entre elles
- On garde tous les prédicteurs

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	4.4209	73.68	73.68	**********
2	0.8561	14.27	87.95	**********
3	0.3731	6.22	94.17	*******
4	0.2139	3.57	97.73	****
5	0.0928	1.55	99.28	**
6	0.0433	0.72	100.00	*

VARIABLES	CORRELATIONS VARIABLE-FACTEUR				
IDEN - LIBELLE COURT	1	2	3	4	5
Cyli - cylindree	0.89	-0.11	0.22	-0.37	-0.05
Puis - puissance	0.89	-0.38	0.11	0.17	0.09
Long - longueur	0.89	0.38	-0.04	0.13	-0.22
Larg - largeur	0.81	0.41	-0.37	-0.10	0.15
Poid - poids	0.91	0.22	0.30	0.14	0.09
Vite - vitesse	0.75	-0.57	-0.30	0.03	-0.06


Représentation des variables quantitatives dans le premier plan factoriel
Facteur 2
Facteur 1
0.8
0.4
0
-0.4
-0.8
-0.8
-0.4
0
0.4
0.8
largeur
longueur
poids
prix
cylindree
puissance
vitesse


Facteur 2 - 14.27 %
1.50
0.75
0
-0.75
-1.50
AUDI-100-L
RENAULT-16-TL
RANCHO
SIMCA-1307-GLS
PEUGEOT-504
PRINCESS-1800-HL
DATSUN-200L
LANCIA-BETA-1300
LADA-1300
Finition=M
Finition=TB
Finition=B
CITROEN-GS-CLUB
TOYOTA-COROLLA
OPEL-REKORD-L
MAZDA-9295
TAUNUS-2000-GL
FIAT-132-1600GLS
RENAULT-30-TS
ALFETTA-1.66
ALFASUD-TI-1350
-3.0
-1.5
0
1.5
3.0
Facteur 1 - 73.68 %

Changement de base

- Formule de reconstitution X=CU'
- Estimation

$$\hat{\beta} = \left(UC'CU'\right)^{-1} UC'y = \left(U\frac{1}{n}C'CU'\right)^{-1}\frac{1}{n}UC'y$$

$$= \left(U\Lambda U'\right)^{-1}\frac{1}{n}UC'y = U\Lambda^{-1}U'U\frac{1}{n}C'y = U\Lambda^{-1}\frac{1}{n}C'y = U\hat{\alpha}$$

- $\hat{\alpha}$ coefficients de régression de y sur les composantes principales

$$\hat{\beta} = U\hat{\alpha} \qquad \hat{\alpha} = U'\hat{\beta} \qquad V(\hat{\alpha}) = \sigma^2\Lambda^{-1}$$

$$V(\hat{\beta}) = \sigma^2 U\Lambda^{-1}U' \qquad V(\hat{\beta}_j) = \sigma^2\sum_{k=1}^{p}\frac{u_{jk}^2}{\lambda_k}$$

prédiction

- Combinaison linéaire des composantes principales

$$\hat{y} = X\hat{\beta} = XU\hat{\alpha} = C\hat{\alpha} = \alpha_1 \mathbf{c}_1 + \ldots + \alpha_p \mathbf{c}_p$$

- Prédicteur approché en éliminant les petites valeurs propres

• *résultats ordonnés selon le nombre de composantes principales conservées :*

dim	RMSE	Intercept	CYL	PUIS	LON	LAR	POIDS	VITESSE
1	4301.68	-43286.46	2.74369	49.978	46.0278	175.804	7.5893	71.383
2	4401.15	-34893.04	2.94823	62.544	34.5556	124.103	6.4980	102.827
3	4451.25	-5360.02	4.31052	75.618	30.1484	-39.880	11.5931	45.222
4	4296.24	-5829.58	-2.62099	131.959	70.7514	-167.635	18.6615	64.667
5	4294.23	-9856.87	-4.01533	181.544	-42.9173	141.908	26.3105	11.216
6	4406.23	-8239.36	-3.50518	282.169	-15.0377	208.694	12.5747	-111.114

• *coefficients de corrélation entre la variable prix et les 6 composantes principales :*

	CORRELATIONS VARIABLE-FACTEUR					
	1	2	3	4	5	6
PRIX	-0.77	0.09	-0.13	-0.23	-0.16	-0.10

L'ordre des corrélations n'est pas celui des valeurs propres

6 Régression PLS

- Proposée par H.et S.Wold, (cf.M.Tenenhaus 1998) proche de la régression sur composantes principales
- projection sur des combinaisons linéaires des prédicteurs non corrélées entre elles
- différence essentielle: composantes PLS optimisées pour être prédictives de Y, alors que les composantes principales ne font qu'extraire le maximum de variance des prédicteurs sans tenir compte de Y.

Régression PLS (2)

- t=Xw
- critère de Tucker

$$\max \mathrm{cov}^2(y ; Xw)$$

- compromis entre maximiser la corrélation entre t_1 et y (régression ordinaire) et maximiser la variance de t_1 (ACP des prédicteurs) :

$$\mathrm{cov}^2(y ; Xw) = r^2(y ; Xw)\, V(Xw)\, V(y)$$

Régression PLS (3)

- solution : w_{1j} proportionnels aux covariances cov(y ; x_j) : coefficients du même signe que les corrélations simples entre y et les x_j ; pas de signes surprenants.
- régression PLS avec une composante $y=c_1t_1+ y_1$
- deuxième composante PLS t_2 en itérant le procédé : régression de y_1 sur les résidus des régressions des x_j avec t_1 puis $y=c_1t_1+ c_2t_2 +y_2$ etc.

Régression PLS (4)

- nombre de composantes PLS choisi par validation croisée.
- la première composante PLS est toujours plus corrélée avec Y que la première composante principale :

$$\mathrm{cov}(y;t_1) = r(y;t_1)\sigma(t_1)\sigma(y) \geq \mathrm{cov}(y;c_1) = r(y;c_1)\sigma(c_1)\sigma(y)$$

$$\text{donc } r(y;t_1)\sigma(t_1) \geq r(y;c_1)\sigma(c_1)$$

$$\sigma(c_1) \geq \sigma(t_1) \text{ d'où } r(y;t_1) \geq r(y;c_1)$$

Régression PLS (5)

- Avantage de la régression PLS : simplicité de son algorithme. Ni inversion, ni diagonalisation de matrices, mais seulement une succession de régressions simples, autrement dit des calculs de produits scalaires. On peut donc traiter de très grands ensembles de données.
- la régression PLS donne en pratique d'excellentes prévisions, même dans le cas d'un petit nombre d'observations et d'un grand nombre de variables.

The PLS Procedure

Percent Variation Accounted for
by Partial Least Squares Factors

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	73.6230	73.6230	60.8374	60.8374

Parameter Estimates for Centered and Scaled Data

	PRIX
Intercept	0.0000000000
CYL	0.1457852413
PUIS	0.1823397520
LON	0.1469668392
LAR	0.1247976334
POIDS	0.1719738622
VITESSE	0.1328131564

Parameter Estimates

	PRIX
Intercept	-39940.36629
CYL	2.56208
PUIS	58.80660
LON	43.68699
LAR	154.34048
POIDS	8.25174
VITESSE	71.89164

Root MSE 4239.06107

Regression PLS 23:21 Wednesday, September 22, 2004

Obs	NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE	NAT	FINITION	PRIX	py
1	ALFASUD-TI-1350	1350	79	393	161	870	165	I	B	30570	29223.10
2	AUDI-100-L	1588	85	468	177	1110	160	D	TB	39990	37552.65
3	SIMCA-1307-GLS	1294	68	424	168	1050	152	F	M	29600	31418.15
4	CITROEN-GS-CLUB	1222	59	412	161	930	151	F	M	28250	28037.70
5	FIAT-132-1600GLS	1585	98	439	164	1105	165	I	B	34900	35354.30
6	LANCIA-BETA-1300	1297	82	429	169	1080	160	I	TB	35480	33444.59
7	PEUGEOT-504	1796	79	449	169	1160	154	F	B	32300	35649.18
8	RENAULT-16-TL	1565	55	424	163	1010	140	F	B	32000	29383.52
9	RENAULT-30-TS	2664	128	452	173	1320	180	F	TB	47700	44692.48
10	TOYOTA-COROLLA	1166	55	399	157	815	140	J	M	26540	24733.94
11	ALFETTA-1.66	1570	109	428	162	1060	175	I	TB	42395	35521.09
12	PRINCESS-1800-HL	1798	82	445	172	1160	158	GB	B	33990	36406.57
13	DATSUN-200L	1998	115	469	169	1370	160	J	TB	43980	41321.71
14	TAUNUS-2000-GL	1993	98	438	170	1080	167	D	B	35010	37219.47
15	RANCHO	1442	80	431	166	1129	144	F	TB	39450	32576.90
16	MAZDA-9295	1769	83	440	165	1095	165	J	M	27900	35059.13
17	OPEL-REKORD-L	1979	100	459	173	1120	173	D	B	32700	39443.09
18	LADA-1300	1294	68	404	161	955	140	U	M	22100	27817.42

7 La régression ridge

- Hoerl et Kennard (1970)

$$b_R = \left(X'X + kI \right)^{-1} X'y$$

- Trois interprétations
  - Estimateur rétréci d'erreur minimale
  - Coefficients bornés
  - Approche bayesienne

7.1 Diminution de l'erreur quadratique

- En régression simple sur données centrées $y=\alpha+\beta x+e$

$$\hat{\alpha}=\bar{y} \quad \text{et} \quad \hat{\beta}=\frac{\sum x_i y_i}{\sum x_i^2}$$

- Estimateur rétréci $\hat{\beta}_R = c\hat{\beta}$

$$E\left(\left(c\hat{\beta}-\beta\right)^2\right)=V\left(c\hat{\beta}\right)+\left(E\left(c\hat{\beta}\right)-\beta\right)^2=c^2V\left(\hat{\beta}\right)+\left(c-1\right)^2\beta^2=c^2\frac{\sigma^2}{\sum x_i^2}+\left(c-1\right)^2\beta^2$$

- Minimum pour $c=\dfrac{\beta^2}{\beta^2+\dfrac{\sigma^2}{\sum x_i^2}}$

- Ridge avec $\hat{\beta}_R = \dfrac{\sum x_i y_i}{\sum x_i^2 + \dfrac{\sigma^2}{\beta^2}}$
- $k = \dfrac{\sigma^2}{\beta^2}$

7.2 Régression à coefficients bornés

- Minimisation de $\|y - Xb\|^2$ sous $\|b\|^2 \leq c^2$

- régularise la solution pour éviter des coefficients instables

7.3 Régression bayesienne

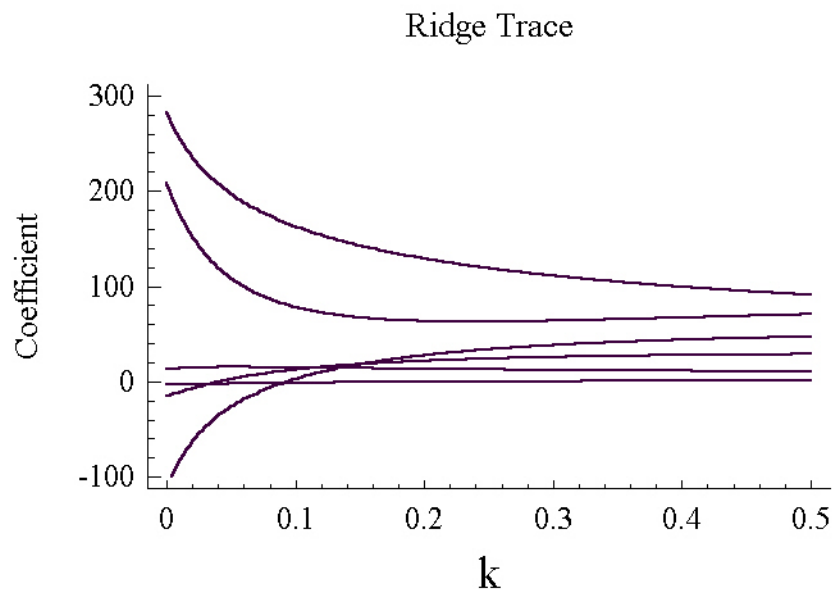
- distribution *a priori* gaussienne sur β $N(0;\psi^2 I)$
- Y/β est une gaussienne $N(X\beta;\sigma^2 I)$
- loi *a posteriori* de β/Y gaussienne
- valeur la plus probable *a posteriori*, espérance *a posteriori* :

$$\hat{\beta} = \left(X'X + \frac{\sigma^2}{\psi^2} I \right)^{-1} X'y$$

Evolution des coefficients de régression en fonction du paramètre *k*.

Regression Coefficients

Ridge Parameter	cylindree	puissance	longueur	largeur	poids	vitesse
0.0	-3.50518	282.169	-15.0377	208.694	12.5747	-111.114
0.05	-2.18019	197.405	2.76652	108.987	15.2924	-26.2437
0.1	-1.30002	163.095	12.6414	78.4137	14.811	3.09658
0.15	-0.693863	142.962	18.2783	67.2553	14.0478	18.3139
0.2	-0.255884	129.251	21.7857	63.497	13.3264	27.6233
0.25	0.0724271	119.112	24.1123	62.9383	12.6918	33.8481
0.3	0.325527	111.21	25.727	63.8295	12.1402	38.2416
0.35	0.524946	104.817	26.8832	65.3631	11.6592	41.4531
0.4	0.684805	99.501	27.7286	67.1422	11.2366	43.8555
0.45	0.814737	94.9847	28.3541	68.9656	10.8621	45.6797
0.5	0.921532	91.0816	28.819	70.7303	10.5273	47.0767



Choix de k:

- *graphique*
- *validation croisée*

SAS fournit les erreurs standard pour RIDGE et RCP

7.4 nombre équivalent de parametres ou ddl effectif

$$df(k) = Trace\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\right)$$

$$= \sum_{j=1}^{P} \frac{n\lambda_j}{n\lambda_j + k}$$

- de façon générale pour un estimateur linéaire

$$df = Trace(\mathbf{S}) \quad \text{si } \hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

(Hastie et al., 2009)

8 Le LASSO

http://www-stat.stanford.edu/~tibs/lasso.html

The Lasso Page

L1-constrained fitting
for statistics and data mining

The Lasso is a shrinkage and selection method for linear regression.

It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients.

- Critère voisin de la ridge:

$$\|y - Xb\|^2 \quad \text{sous} \sum_{j=1}^{p} |b_j| < c$$

- Pénalité L_1 au lieu de L_2. Pas de solution analytique
- Si c est petit, certains coefficients seront nuls ⇒ sélection
- Si $c > \sum_{j=1}^{p} |b_{jols}|$ on retrouve la régression multiple usuelle

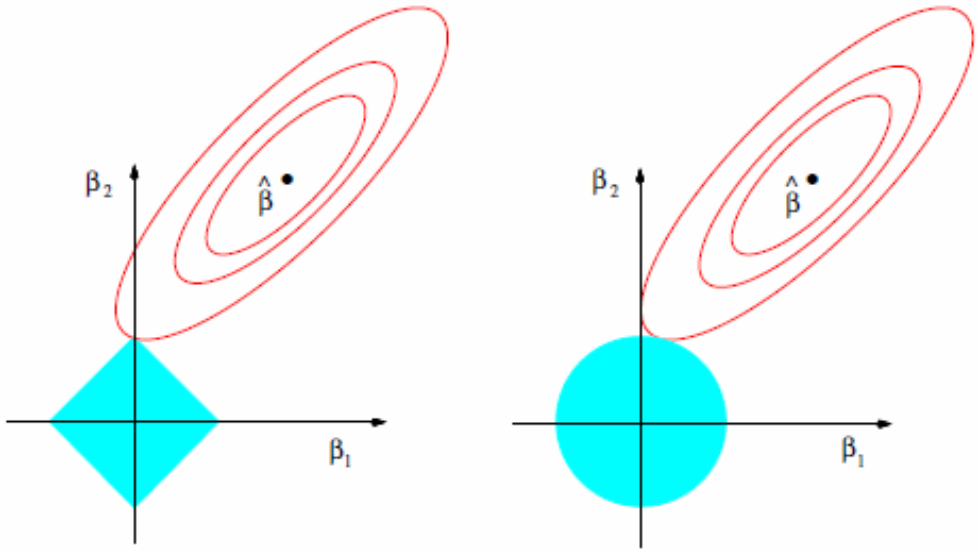


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

Lasso et PROC GLMSELECT

```
ods graphics on;
proc glmselect data=bagnole plots=all;
model prix=cyl puis lon lar poids vitesse /
selection=lasso (stop=7 choose=BIC);
run;
```

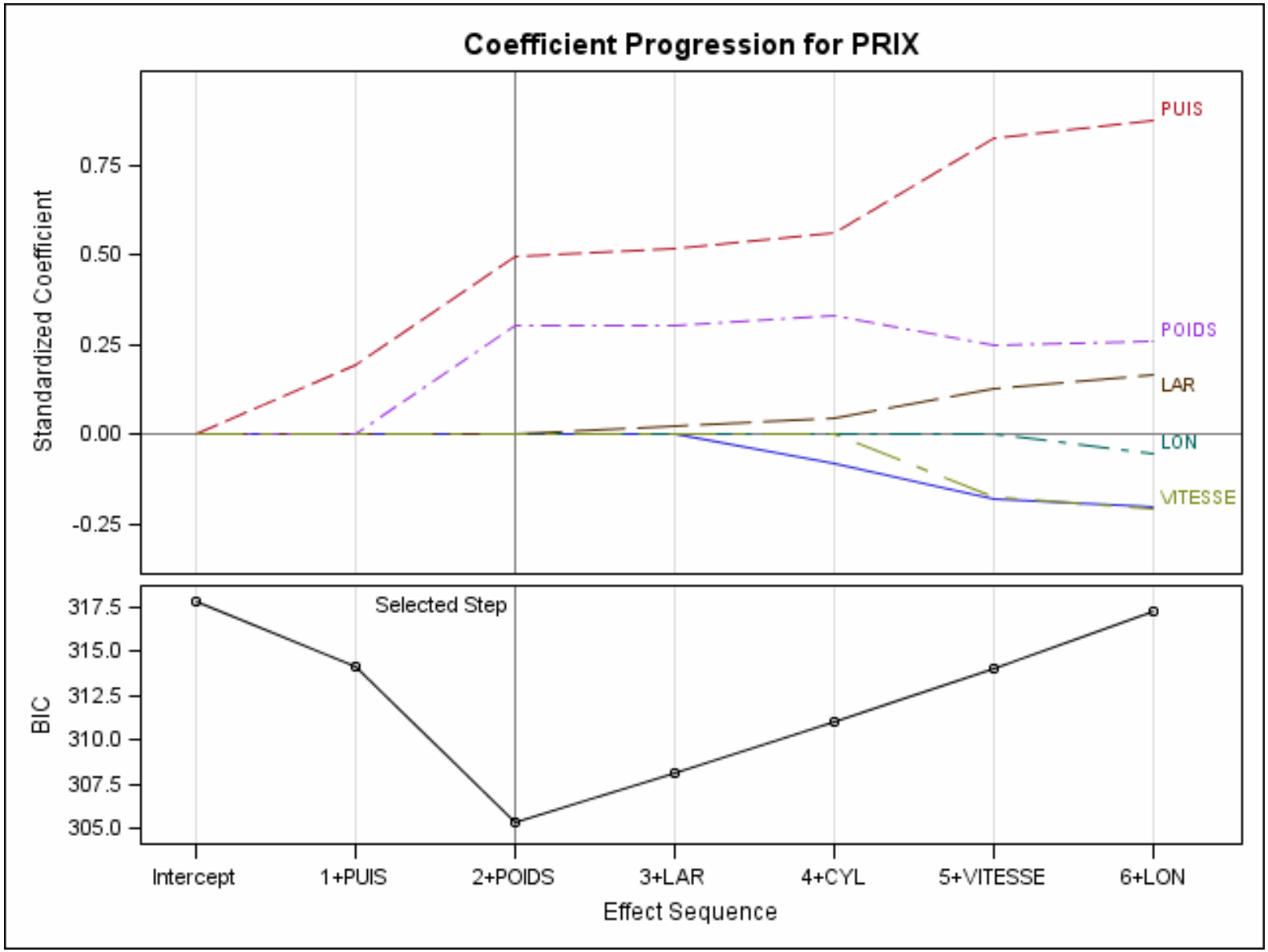
Données pour regression
The GLMSELECT Procedure

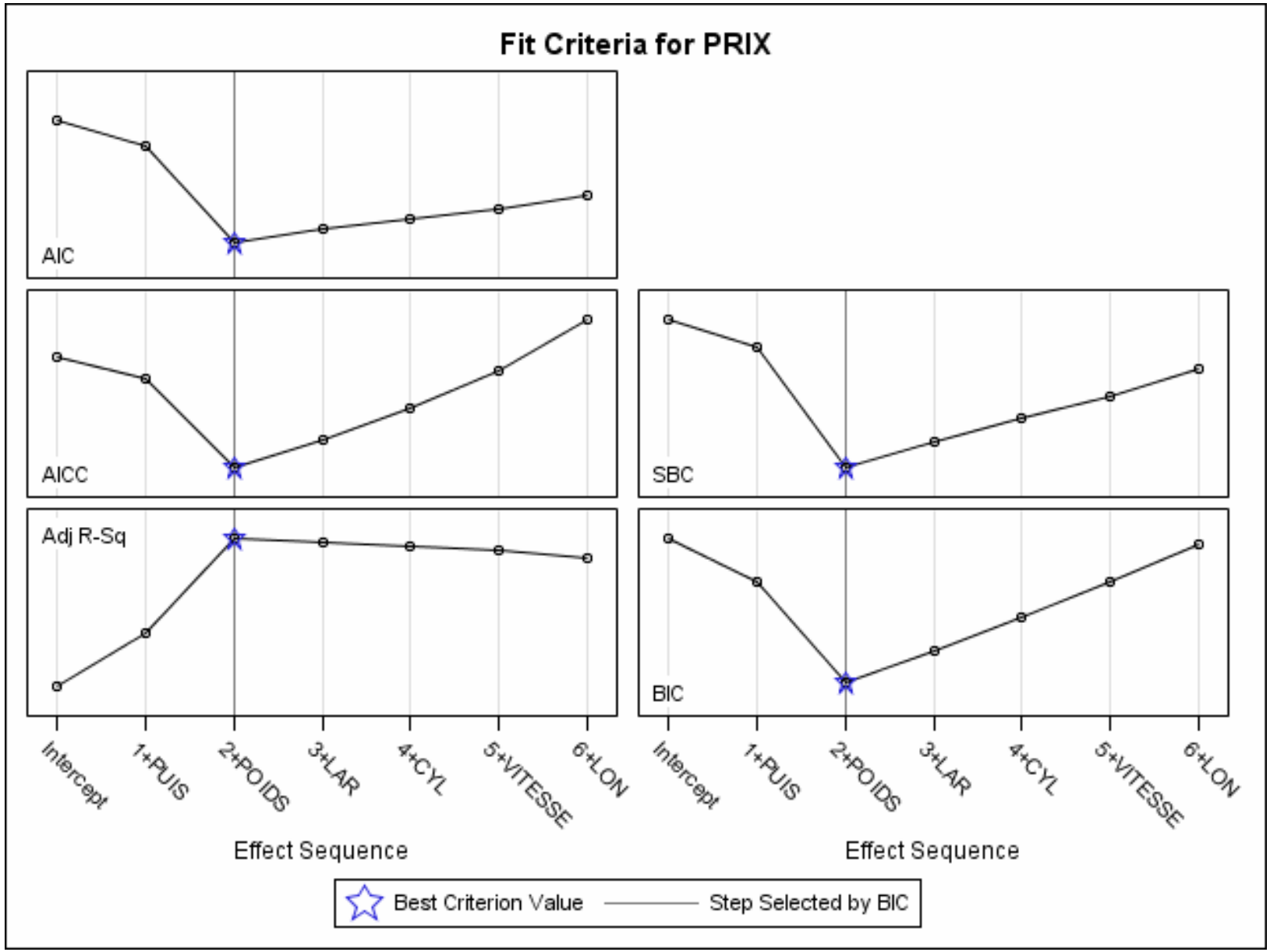
LASSO Selection Summary

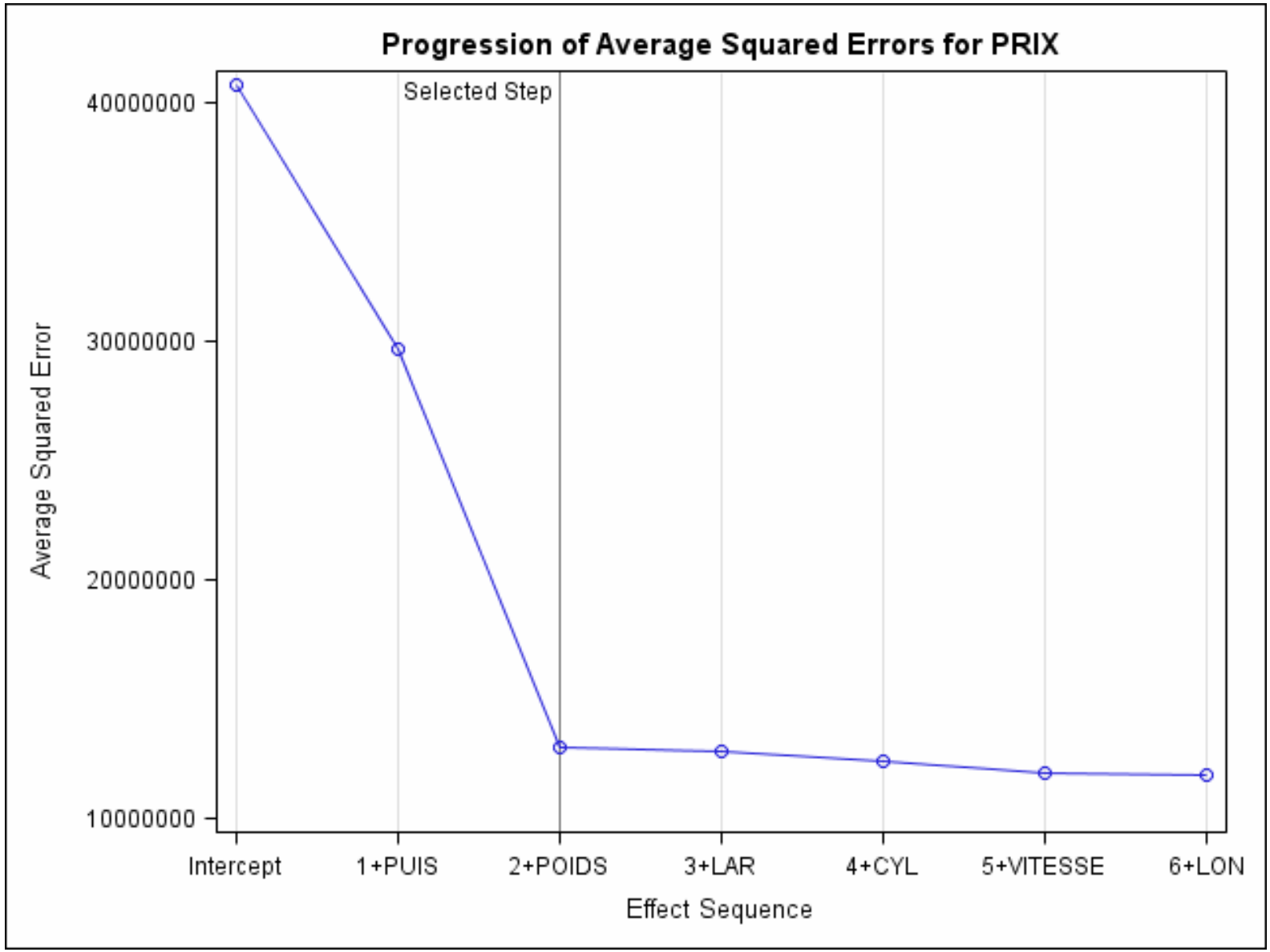
Step	Effect Entered	Effect Removed	Number Effects In	BIC
0	Intercept		1	317.8324
1	PUIS		2	314.0991
2	POIDS		3	305.3416*
3	LAR		4	308.1679
4	CYL		5	311.0317
5	VITESSE		6	314.0364
6	LON		7	317.3025

* Optimal Value Of Criterion

Selection stopped because all effects are in the final model.


Coefficient Progression for PRIX
Standardized Coefficient
0.75
0.50
0.25
0.00
-0.25
PUIS
POIDS
LAR
LON
VITESSE
BIC
317.5
315.0
312.5
310.0
307.5
305.0
Selected Step
Intercept
1+PUIS
2+POIDS
3+LAR
4+CYL
5+VITESSE
6+LON
Effect Sequence


Fit Criteria for PRIX
AIC
AICC
SBC
Adj R-Sq
BIC
Intercept
1+PUIS
2+POIDS
3+LAR
4+CYL
5+VITESSE
6+LON
Effect Sequence
Effect Sequence
Best Criterion Value
Step Selected by BIC


Progression of Average Squared Errors for PRIX
Selected Step
40000000
30000000
20000000
10000000
Average Squared Error
Intercept
1+PUIS
2+POIDS
3+LAR
4+CYL
5+VITESSE
6+LON
Effect Sequence

The GLMSELECT Procedure
Selected Model

The selected model, based on BIC, is the model at Step 2.

Effects: Intercept PUIS POIDS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	499772877	249886439	15.99
Error	15	234382913	15625528	
Corrected Total	17	734155790		

Root MSE	3952.91380
Dependent Mean	34159
R-Square	0.6807
Adj R-Sq	0.6382
AIC	320.87771
AICC	323.95463
BIC	305.34160
C(p)	0.07232
SBC	303.54882

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	5002.288413
PUIS	1	159.803389
POIDS	1	14.492675

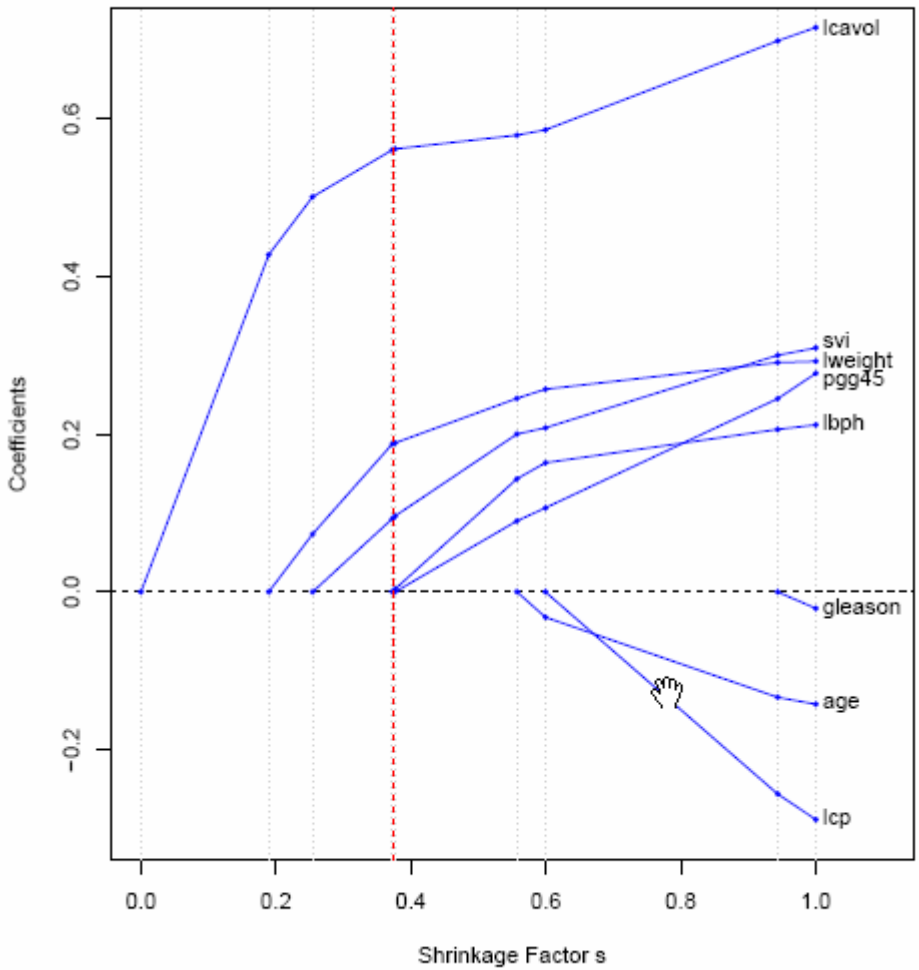


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 9; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed;*

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 3

- Interprétation bayesienne:
  - loi a priori de Laplace ou double exponentielle sur chaque β_j

$$f(\beta_j) = \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau} \right)$$

- Estimateur non linéaire

- Avantages et inconvénients
- Pour
  - Le Lasso rétrécit les coefficients vers zéro de façon continue.
  - Produit un modèle parcimonieux.
  - Est une méthode de sélection.
- Contre
  - le nombre de variables sélectionnées est limité par n
  - Inadapté au cas des puces à ADN n(arrays)<<p(genes)
  - Choisit une seule variable dans un groupe de variables très corrélée

Une variante « lasso hybrid selection » pour obliger SAS à faire de la validation croisée avec le critère PRESS

```
proc glmselect data=bagnole plots=all;
model prix=cyl puis lon lar poids vitesse /
 selection=lasso  (stop=7 lscoeffs choose=Press);
run;
```

LSCOEFFS
requests a hybrid version of the LAR and LASSO methods, where the sequence of models is determined by the LAR or LASSO algorithm but the coefficients of the parameters for the model at any step are determined by using ordinary least squares.

The GLMSELECT Procedure
Selected Model

The selected model, based on PRESS, is the model at Step 2.

Effects: Intercept PUIS POIDS

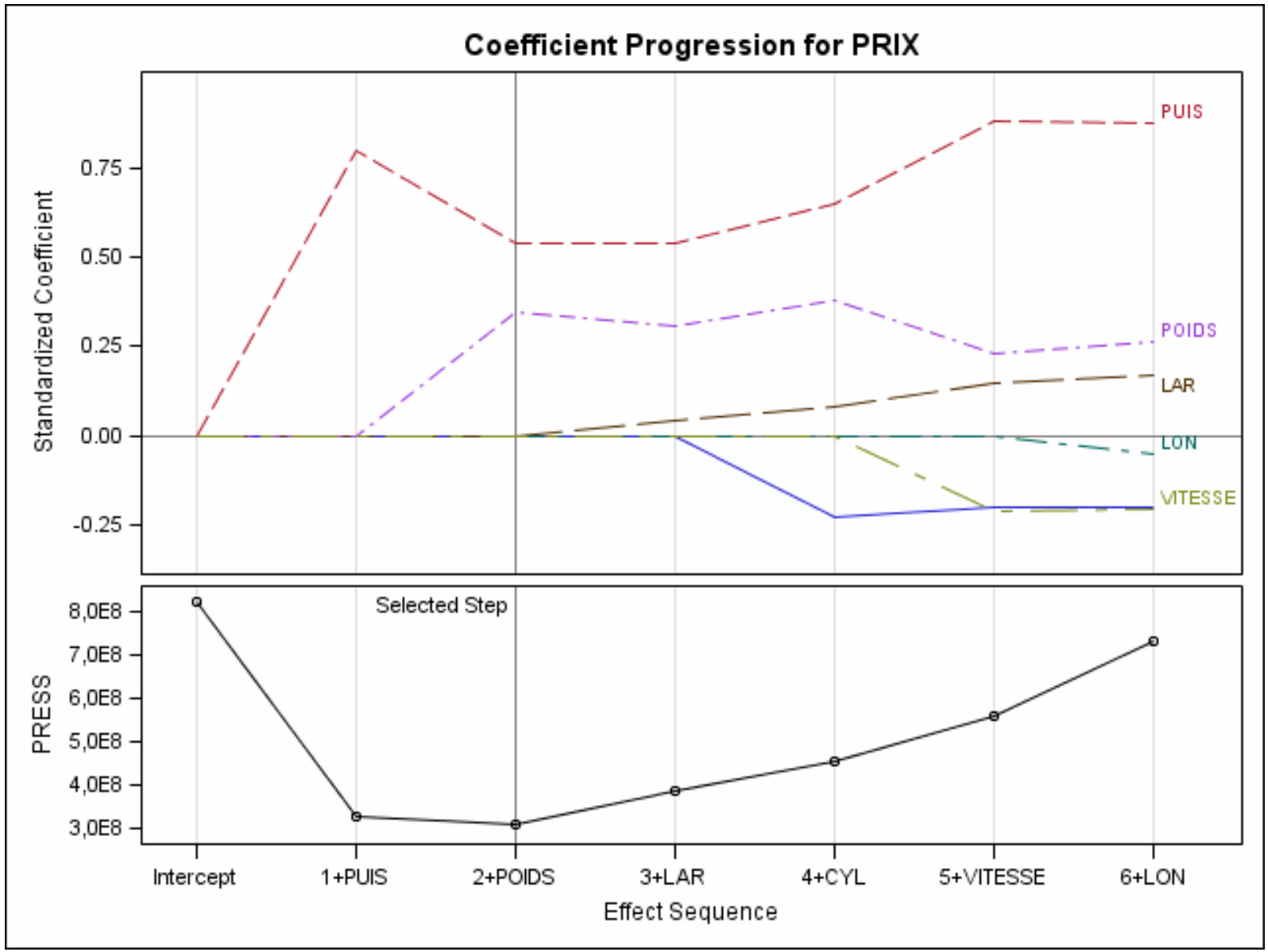
Analysis of Variance

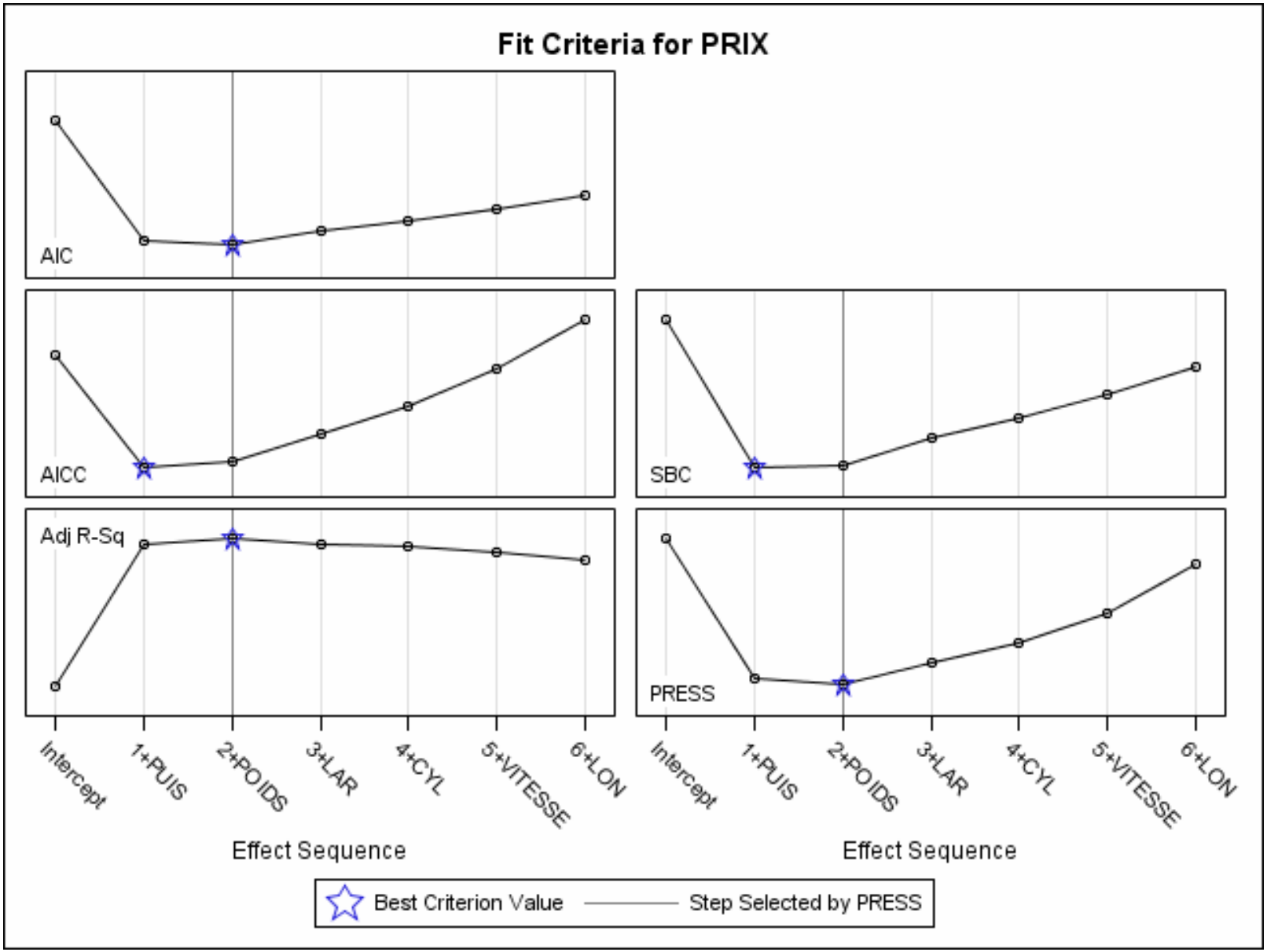
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	504091154	252045577	16.43
Error	15	230064636	15337642	
Corrected Total	17	734155790		

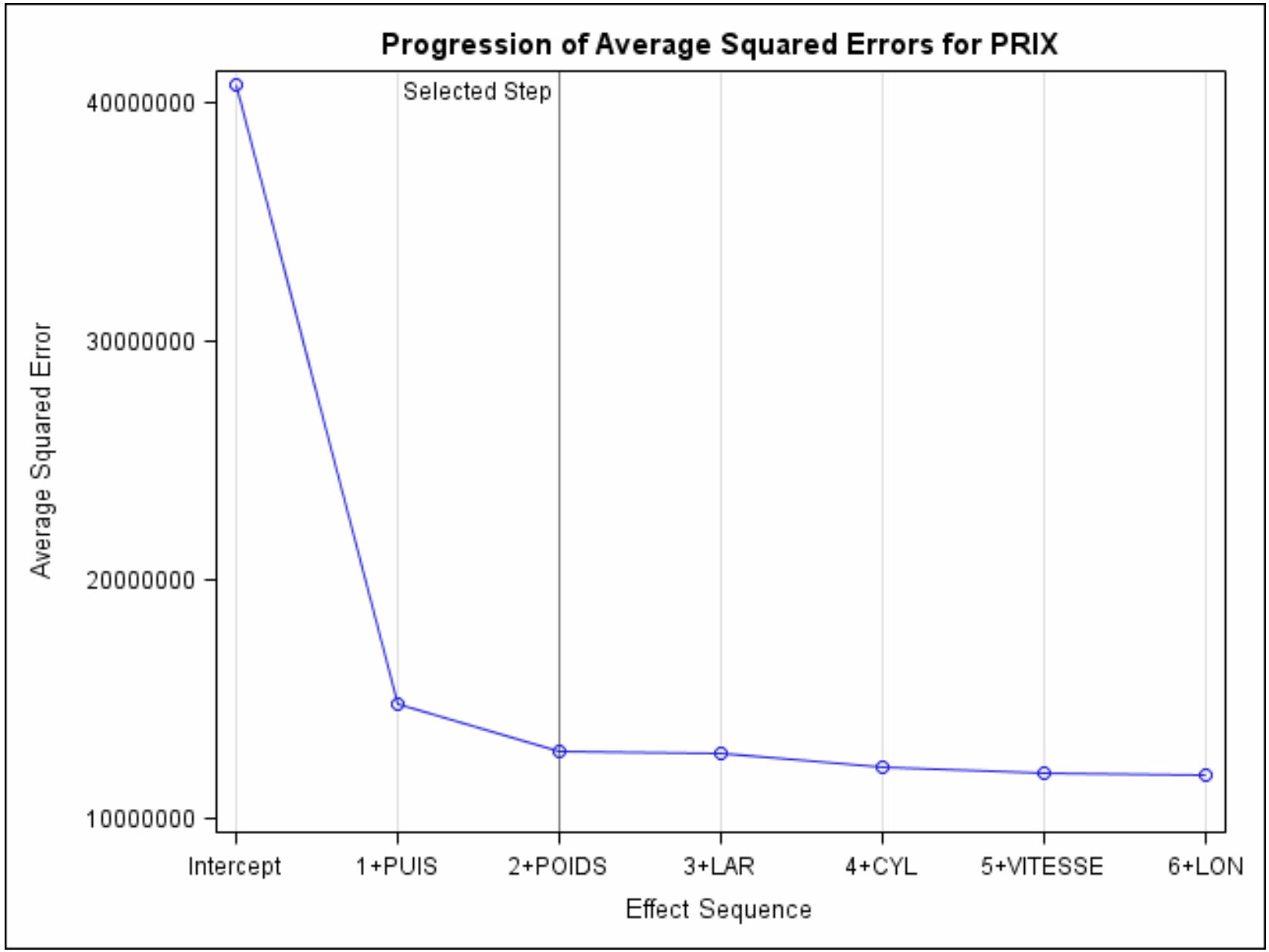
Root MSE	3916.33023
Dependent Mean	34159
R-Square	0.6866
Adj R-Sq	0.6448
AIC	320.54298
AICC	323.61991
PRESS	308496438
SBC	303.21410

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	1775.601201
PUIS	1	172.967225
POIDS	1	16.451161


Coefficient Progression for PRIX
Standardized Coefficient
0.75
0.50
0.25
0.00
-0.25
PUIS
POIDS
LAR
LON
VITESSE
Selected Step
PRESS
8,0E8
7,0E8
6,0E8
5,0E8
4,0E8
3,0E8
Intercept
1+PUIS
2+POIDS
3+LAR
4+CYL
5+VITESSE
6+LON
Effect Sequence


Fit Criteria for PRIX
AIC
AICC
SBC
Adj R-Sq
PRESS
Intercept
1+PUIS
2+POIDS
3+LAR
4+CYL
5+VITESSE
6+LON
Effect Sequence
Effect Sequence
Best Criterion Value
Step Selected by PRESS


Progression of Average Squared Errors for PRIX
Selected Step
40000000
30000000
20000000
10000000
Average Squared Error
Intercept
1+PUIS
2+POIDS
3+LAR
4+CYL
5+VITESSE
6+LON
Effect Sequence

9. Elastic net

- Combine les pénalités de la ridge et du lasso

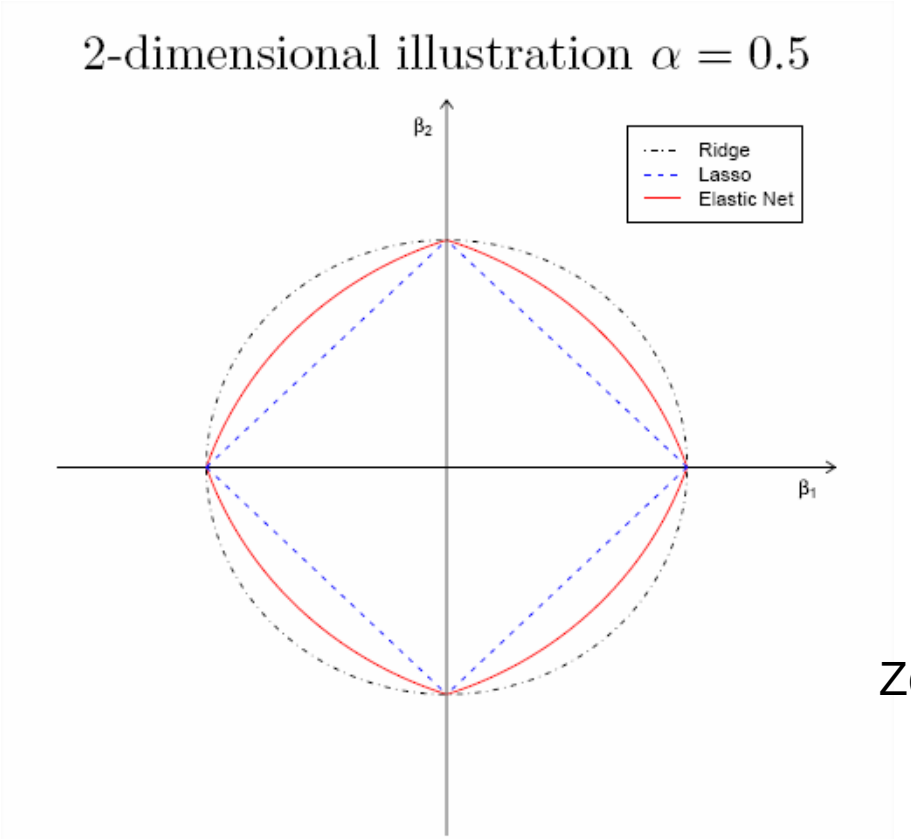
$$\min\left(\|y - Xb\|^2 + \lambda_2 \|b\|^2 + \lambda_1 \|b\|_1\right)$$

- autre formulation:

$$\min \|y - Xb\|^2 + \lambda \sum_{j=1}^{p} \left(\alpha b_j^2 + (1-\alpha)|b_j|\right)$$

$$avec \quad \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

- La partie L_1 conduit à un modèle «sparse»
- La partie L_2 enlève la limitation sur le nombre de variables retenues et favorise le choix de groupes



Zou et Hastie

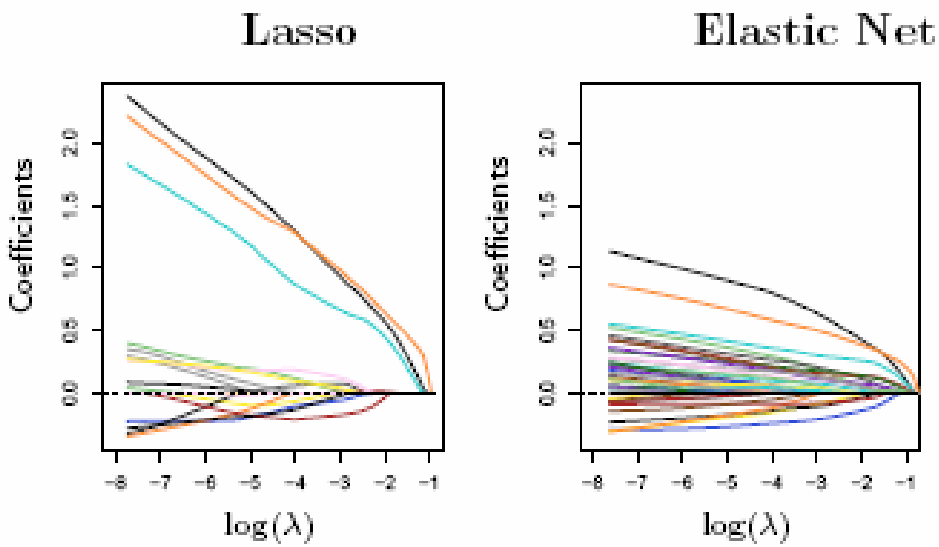


FIGURE 18.5. *Regularized logistic regression paths for the leukemia data. The left panel is the lasso path, the right panel the elastic-net path with* $\alpha = 0.8$. *At the ends of the path (extreme left), there are* 19 *nonzero coefficients for the lasso, and* 39 *for the elastic net. The averaging effect of the elastic net results in more non-zero coefficients than the lasso, but with smaller magnitudes.*

Références

- Birkes D. , Dodge Y. (2003) *Alternative methods of regression*, Wiley
- Hastie T., Tibshirani R., Friedman J. (2009) *The elements of statistical learning*, 2nd edition, Springer, 2009
- Tenenhaus M. (1998) *La régression PLS*, Technip
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J*. Royal. Statist. Soc B.,* Vol. 58, No. 1, pages 267-288).
- Weisberg S. (1980) *Applied linear regression*, Wiley