



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΕΛΕΓΧΟΣ ΤΗΣ ΔΟΜΙΚΗΣ ΔΙΑΣΤΑΣΗΣ ΠΡΟΒΛΗΜΑΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ρεβέκκα Α. Χριστοπούλου

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος

Συμπληρωματικής Ειδίκευσης στη Στατιστική

Μερικής Παρακολούθησης (Part-time)

Αθήνα
Φεβρουάριος 2007



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ



0 000000 595438





ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
εισ 80972
Αρ.
παξ.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Έλεγχος της Δομικής Διάστασης Προβλημάτων Παλινδρόμησης

Ρεβέκκα Α. Χριστοπούλου

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος

Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

Αθήνα
Φεβρουάριος 2007





ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
εισ80972
Αρ.
ταξ.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Εργασία που υποβλήθηκε ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

ΕΛΕΓΧΟΣ ΤΗΣ ΔΟΜΙΚΗΣ ΔΙΑΣΤΑΣΗΣ ΠΡΟΒΛΗΜΑΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ρεβέκκα Α. Χριστοπούλου

Υπεύθυνο μέλος ΔΕΠ:
Π. Τσιαμυρτζής
Λέκτορας

Ο Διευθυντής Μεταπτυχιακών Σπουδών

Επαμεινώνδας Πανάς
Καθηγητής



ΑΦΙΕΡΩΣΗ

Θα ήθελα να αφιερωστώ τον καθηγητή μου κ. Π. Γεωργοπόύλη για τη
ριζήβια του στη διδακτηριακή μου εργασία.

**Στο σύζυγό μου Παναγιώτη Κανελάτο, το γιο μου Κωνσταντίνο και την
κόρη μου Αριστέα**



Το αξέτιλο πανεπιστήμιο για την Καραϊσκάκη που είδα
εγώ την γεννάδα



ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΣ ΕΥΧΑΡΙΣΤΙΕΣ

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Π. Τσιαμυρτζή για τη βοήθειά του στη διπλωματική μου εργασία.

(Πανεπιστημιούπολη, από την πανεπιστημιακή του έτοιμη)

Εργάσθηκε στην Nestle Hellas από το 1975 έως το 2014 ως τεχνικός Logistics and Supply Chain Coordinator. Οι αρμόδιες του ήταν: Προβίβαση των επιχειρήσεων σε απότομη Ηπειρωτικότητα, παραγάν. Αντιλήψη Αποθήκης, Διαχείριση αποθέματων (από το Customer Service Level, Υπεύθυνη Επειγόντων / Εξαγωγών).

Το Δεκέμβριο του 2003 πήρε επίδειξη παραγάν. ρα θέμα: "Μαρτυρεία Απόθεκης και Σύγκριση Αντιλήψης" στη σημερινή η εταιρία International Business Center Hellas.

Το ίδιο έτος τον 2004 πήρε επίδειξη παραγάν. ρα θέμα: "Ο Έλλος των Logistics και Συγκεκριμένη Εργασία της Nestle, της Unilever" με σχεδιαστή την επιχείρηση Επικοινωνιών.

Το Δεκέμβριο του 2004 πήρε επίδειξη παραγάν. ρα θέμα: "Επειγόντων - Εξαγωγών Συστηματική Ανάπτυξη - Επειγόντων - Εξαγωγών (Ιανουάριος 2005)" με σχεδιαστή την επιχείρηση Επειγόντων.

Το Δεκέμβριο του 2004 πήρε επίδειξη παραγάν. ρα θέμα: "Σύγκριση και Διαπομπή Απόθεκης και απότομης προσφοράς θεωρείσιμων πληρωμών".

Στάθηκε την γεννητική παραγάν. ρα θέμα: "Η Επιχείρηση και Οικονομικό Πλεονεκτήμα της Επιχείρησης".

ΕΥΑΓΓΕΛΙΟ

Θα γίνεται επιδοτητής του καρφιάτη της Η Γαλατούρας
για την απόπειρα να φύγει από την Ελλάδα



ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Γεννήθηκα το 1969 στην Αθήνα. Ολοκλήρωσα τις προπτυχιακές μου σπουδές στο Πανεπιστήμιο Πατρών, στο τμήμα Φυσικής το 1994.

Εργάστηκα στην εταιρεία Energizer από το 1998 έως το 2004 στο τμήμα Logistics ως Supply Chain Coordinator. Οι αρμοδιότητές μου ήταν: Πρόβλεψη και εκτίμηση των πωλήσεων, Προγραμματισμός αναγκών, Αναπλήρωση Αποθεμάτων, Διαχείριση αποθεμάτων με βάση το Customer Service Level, Υπεύθυνη Εισαγωγών / Εξαγωγών.

Το Φεβρουάριο του 2003 παρακολούθησα σεμινάριο με θέμα: "Management Αποθεμάτων και Τεχνικές Αναπλήρωσης" που οργάνωσε η εταιρεία International Training Center Hellas.

Τον Οκτώβριο του 2001 παρακολούθησα σεμινάριο με θέμα: "Ο Ρόλος των Logistics στο Στρατηγικό Σχεδιασμό της Πολιτικής των Εταιρειών" που οργάνωσε η εταιρεία Οικονομοτεχνική.

Το Φεβρουάριο του 2001 παρακολούθησα σεμινάριο με θέμα: "Εισαγωγές - Εξαγωγές Τεχνική & Διαδικασίες Τριγωνικές - Επαναληπτικές (Incoterms 2000)" που οργάνωσε η εταιρεία Οικονομοτεχνική.

Το Νοέμβριο του 1999 παρακολούθησα σεμινάριο με θέμα: "Οργάνωση και Λειτουργία Αποθηκών" που οργάνωσε η εταιρεία Οικονομοτεχνική.

Συνέχισα τις μεταπτυχιακές μου σπουδές στο τμήμα Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών.

BΙΟΛΑΦΙΚΟ ΕΜΕΙΟΥΜΑ

BΙΟΛΑΦΙΚΟ ΕΜΕΙΟΥΜΑ

παντού σε όλη την Ελλάδα. Οργανώσατε μια πανεπιστημιακή διαδικασία για την ανάπτυξη της οικονομίας της Ελλάδας.

Επαγγελματικά στην επαγγελματική παραγωγή από το 1991 έως το 2004 από την Τεχνητή Λογιστική της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών. Η πανεπιστημιακή παραγωγή από την Τεχνητή Λογιστική της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών στην Ελλάδα.

Το Φεβρουάριο του 2003 παρακολούθησα σε πανεπιστημιακό επίπεδο την θέση: "Μαναγερός εργοθήκης" στην Κατοικητική Σερβισούς Ιανέλ. Υπέβαλλε την Εργασία της στην Επαγγελματική Σχολή της Επαγγελματικής Πανεπιστημίου Αθηνών.

Το Οκτώβριο του 2001 παρακολούθησε σε πανεπιστημιακό επίπεδο την θέση: "Ο Ψήφος των Logistics της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών". Η πανεπιστημιακή παραγωγή από την Τεχνητή Λογιστική της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών.

Το Φεβρουάριο του 1993 παρακολούθησε σε πανεπιστημιακό επίπεδο την θέση: "Επαγγελματικός Διευθυντής Στοχεύοντας στην Επαγγελματική Σχολή της Επαγγελματικής Πανεπιστημίου Αθηνών". Η πανεπιστημιακή παραγωγή από την Τεχνητή Λογιστική της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών.

Το Νοέμβριο του 1992 παρακολούθησε σε πανεπιστημιακό επίπεδο την θέση: "Οργανωτής και Βετούδης της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών". Η πανεπιστημιακή παραγωγή από την Τεχνητή Λογιστική της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών.

Συνέβησα την παρακολούθηση την πανεπιστημιακή παραγωγή από την Τεχνητή Λογιστική της Επαγγελματικής Σχολής της Επαγγελματικής Πανεπιστημίου Αθηνών.

ABSTRACT

Διάφορα Χρονικά έργα

Rebecca Christopoulou

Επίκληση της Δομής Λεύκων: Απόστολος Προβλημάτων

Discovering the structural dimension in regression problems

February 2007

One of the primary objects in a regression analysis is to understand how the response variable depends on one or more predictors. Things are easy when we have only one predictor, as a two dimensional plot is necessary to observe the dependence between the two variables. However it is more difficult when we have more than one predictors. In this dissertation, is about how to use graphs in order to understand how a response variable depends on one or more predictors. We use 2D scatterplots, 3D scatterplots and scatterplot matrices. We also describe the way to discover regressions' structural dimension through graphs and arithmetic procedures. We finally present the notions and methods through the analysis of two problems.



ABSTRACT

Reveccs Cryptobonon

Discovering the structural dimension in regression biplotts

February 2002

One of the primary objects in a regression analysis is to understand how the response variable depends on one or more predictors. Little else can we base only one predictor, as a two dimensional plot is necessary to observe the dependence between the two variables. However, it is more difficult when we have more than one predictor. In this dissertation, a good way to see relationships in order to understand how a response variable depends on one or more predictors. We use 2D scatterplots and 3D scatterplots and scatterplot matrices. We also describe the way to discover relationships, structural dimension through biplots and biplot matrices. We finally present the notions and methods concerning the analysis of two biplots.

ΠΕΡΙΛΗΨΗ

Ρεβέκκα Χριστοπούλου

Έλεγχος της Δομικής Διάστασης Προβλημάτων Παλινδρόμησης

Φεβρουάριος 2007

Ο πρωταρχικός σκοπός σε μια ανάλυση παλινδρόμησης, είναι να κατανοήσουμε τον τρόπο με τον οποίο μια εξαρτημένη μεταβλητή (response variable) εξαρτάται από μια ή περισσότερες ανεξάρτητες ή ελεγχόμενες (predictors) τυχαίες μεταβλητές. Στην περίπτωση που έχουμε μόνο μια ανεξάρτητη μεταβλητή, τότε αρκεί ένα διάγραμμα δυο διαστάσεων για να παρατηρήσουμε την εξάρτηση μεταξύ των δυο μεταβλητών. Τα πράγματα όμως γίνονται δυσκολότερα όταν έχουμε περισσότερες από μια ανεξάρτητες μεταβλητές. Στη διπλωματική αυτή, ασχολούμαστε με το πώς μπορούμε να χρησιμοποιούμε διαγράμματα για να κατανοούμε πώς η εξαρτημένη μεταβλητή εξαρτάται από μια ή περισσότερες ανεξάρτητες μεταβλητές. Χρησιμοποιούμε διδιάστατα και τρισδιάστατα διαγράμματα διασποράς καθώς και πίνακες διαγραμμάτων διασποράς. Επίσης περιγράφουμε τον τρόπο με τον οποίο μπορούμε να ανακαλύπτουμε τη δομική διάσταση προβλημάτων παλινδρόμησης μέσω διαγραμμάτων και αριθμητικών μεθόδων. Τέλος παρουσιάζουμε τις έννοιες και μεθόδους μέσω της ανάλυσης δυο εφαρμογών.





ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Ευχαριστίες	V
Βιογραφικό σημείωμα	VII
Abstract	IX
Περίληψη	XI
Κεφάλαιο 1:	1
Γραφικές μέθοδοι παλινδρόμησης	1
1.1 Εισαγωγή	1
1.2 Διαγράμματα απλής παλινδρόμησης	2
1.2.1 Διαγράμματα διασποράς	2
1.2.2 Γραμμική παλινδρόμηση	4
1.2.3 Χρήση των διαγραμμάτων για την αξιολόγηση γραμμικότητας	5
1.3 Διαγράμματα δυο διαστάσεων	6
1.3.1 Πίνακες διαγραμμάτων διασποράς	7
1.3.2 Διαγράμματα μερικής απόκρισης	8
1.4 Διαγράμματα τριών διαστάσεων	9
1.5 Απεικόνιση γραμμικής παλινδρόμησης με δύο ανεξάρτητες μεταβλητές ..	11
1.5.1 Το ιδανικό διάγραμμα περύληψης	12
1.5.2 Προσαρμογή της ευθείας παλινδρόμησης	12
1.5.3 Κατανομή των ανεξάρτητων μεταβλητών	14
Κεφάλαιο 2:	15
Έλεγχος της δομικής διάστασης της παλινδρόμησης	15
2.1 Εισαγωγή	15
2.1.1 Γενικά τρισδιάστατα διαγράμματα απόκρισης	15
2.1.2 Έλεγχος ενός εκτιμημένου διαγράμματος περύληψης	18
2.1.3 Δομική διάσταση παλινδρόμησης με πολλές ανεξάρτητες μεταβλητές ..	20
2.2 Εύρεση της διάστασης της παλινδρόμησης	24
2.2.1 Γραφική μέθοδος εύρεσης των διαστάσεων	24
2.3 Τμηματική αντίστροφη παλινδρόμηση	30
2.4 Μέθοδος SAVE	38
2.5 Principal Hessian directions	43
2.6 Συμπεράσματα	44
Κεφάλαιο 3:	47
Εφαρμογές	47
3.1 Εισαγωγή	47
3.2 Ανάλυση προβλήματος Smoking and Cancer	47
3.2.1 Περιγραφή του προβλήματος	47
3.2.2 Διδιάστατη γραφική απεικόνιση των δεδομένων	48
3.2.3 Τρισδιάστατη γραφική απεικόνιση	54
3.2.4 Απεικόνιση της γραμμικής παλινδρόμησης	56
3.2.5. Εύρεση της δομής του προβλήματος	60
3.3 Ανάλυση προβλήματος Ais	64
3.3.1 Περιγραφή του προβλήματος	64
3.3.2 Διδιάστατη γραφική απεικόνιση των δεδομένων	65
3.3.3 Τρισδιάστατη γραφική απεικόνιση	68
3.3.4 Απεικόνιση της γραμμικής παλινδρόμησης	70
3.3.5. Εύρεση της δομής του προβλήματος	74

Κεφάλαιο 4:	81
Συμπεράσματα	81
4.1 Συνοπτική παρουσίαση	81
4.1.1 Γραφική μέθοδος	81
4.1.2 Αριθμητικές μέθοδοι	82
4.2 Κριτική θεώρηση των μεθόδων	84
Βιβλιογραφία	87

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452		

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

<u>Πίνακας</u>	<u>Σελίδα</u>
Πίνακας 2.1: Περιγραφή μεταβλητών προβλήματος BigMac	27
Πίνακας 2.2: Αποτελέσματα μεθόδου <i>SIR</i>	34
Πίνακας 2.3: Αποτελέσματα μεθόδου <i>SIR</i> με 15 slices	37
Πίνακας 2.4: Αποτελέσματα μεθόδου <i>SIR</i> με 15 slices και εξαρτημένη μεταβλητή την $\log(BigMac)$	38
Πίνακας 3.1: Περιγραφή μεταβλητών προβλήματος <i>Smoking and Cancer</i>	48
Πίνακας 3.2: Οι γραμμικοί συνδυασμού που αποτελούν τους άξονες του διαγράμματος	56
Πίνακας 3.3: Οι γραμμικοί συνδυασμού που αποτελούν τους άξονες του διαγράμματος του Σχήματος 3.8	57
Πίνακας 3.4: Αποτελέσματα παλινδρόμησης της <i>Cig</i> την h^*	58
Πίνακας 3.5: Αποτελέσματα μεθόδου <i>SIR</i> με 14slices	63
Πίνακας 3.6: Αποτελέσματα μεθόδου <i>SIR</i> 14 slices και μετασχηματισμένη εξαρτημένη μεταβλητή	64
Πίνακας 3.7: Περιγραφή μεταβλητών προβλήματος <i>Ais</i>	65
Πίνακας 3.8: Οι γραμμικοί συνδυασμού που αποτελούν τους άξονες του τρισδιάστατου διαγράμματος	70
Πίνακας 3.9: Οι γραμμικοί συνδυασμού που αποτελούν τους άξονες του διαγράμματος του Σχήματος 3.19	71
Πίνακας 3.10: Αποτελέσματα παλινδρόμησης της <i>LBM</i> με την h^*	72
Πίνακας 3.11: Αποτελέσματα μεθόδου <i>SIR</i> με 25 slices	78
Πίνακας 3.12: Αποτελέσματα μεθόδου <i>SIR</i> με 25 slices και ανεξάρτητες μεταβλητές τις <i>Wt</i> και <i>Ht</i>	79



ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

	<u>Σελίδα</u>
Σχήμα	3
Σχήμα 1.1: Διάγραμμα διασποράς	3
Σχήμα 1.2: Διάγραμμα διασποράς με slices στις τιμές $x = 170$ και $x = 190$	3
Σχήμα 1.3: Διάγραμμα καταλοίπων	5
Σχήμα 1.4: Το διάγραμμα διασποράς με διαφορετικό aspect ratio	7
Σχήμα 1.5: Πίνακας διαγραμμάτων διασποράς	8
Σχήμα 1.6: Τρισδιάστατο διάγραμμα	9
Σχήμα 1.7: Διάφορες διδιάστατες απεικονίσεις ενός τρισδιάστατου διαγράμματος	10
Σχήμα 2.1: Διάγραμμα περίληψης	20
Σχήμα 2.2: Ασυχέτιστη απεικόνιση	20
Σχήμα 2.3: Τρισδιάστατο διάγραμμα μεταξύ των μεταβλητών <i>BigMac</i> , <i>log(Bread)</i> και <i>log(TeachSal)</i>	28
Σχήμα 2.4: Πίνακας διαγραμμάτων διασποράς με μετασχηματισμένες σε λογαριθμική κλίμακα ανεξάρτητες μεταβλητές	29
Σχήμα 2.5: Διαγράμματα αντίστροφης παλινδρόμησης με προσαρμοσμένη <i>lowess</i>	29
Σχήμα 2.6: Τμηματοποίηση όταν η y είναι τετραγωνική συνάρτηση ενός από τα z	41
Σχήμα 3.1: Πίνακας διαγραμμάτων διασποράς των δεδομένων	49
Σχήμα 3.2: Πίνακας διαγραμμάτων διασποράς μετά τη διαγραφή του απομονωμένου σημείου	50
Σχήμα 3.3: Πίνακας διαγραμμάτων διασποράς μετά τη διαγραφή των δυο επιπλέον απομονωμένων σημείων	51
Σχήμα 3.4: Διάγραμμα διασποράς των μεταβλητών <i>Kid</i> και <i>Blad</i>	52
Σχήμα 3.5: Διάγραμμα διασποράς των μετασχηματισμένων μεταβλητών <i>Kid</i> και <i>Blad</i>	52
Σχήμα 3.6: Διάγραμμα-καταλοίπων με προσαρμοσμένη <i>lowess</i> καμπύλη	53

Σχήμα 3.7: Πίνακας διαγραμμάτων διασποράς με μετασχηματισμένες ανεξάρτητες μεταβλητές	53
Σχήμα 3.8: Τρισδιάστατο διάγραμμα μεταξύ των μεταβλητών <i>Cig</i> , (<i>Blad</i>) ⁴ και (<i>Leuk</i>) ⁴	55
Σχήμα 3.9: Διάφορες απεικονίσεις του περιστραμμένου διαγράμματος	55
Σχήμα 3.10: Η διδιάστατη απεικόνιση με την πιο ισχυρή γραμμική τάση	56
Σχήμα 3.11: Διδιάστατη απεικόνιση { <i>Cig</i> , h_{ols} }	58
Σχήμα 3.12: Διάγραμμα περίληψης και ασυσχέτιστη απεικόνιση αυτού	59
Σχήμα 3.13: Διαγράμματα αντίστροφης παλινδρόμησης με προσαρμοσμένη <i>lowess</i>	61
Σχήμα 3.14: Πίνακας διαγραμμάτων διασποράς των δεδομένων <i>Ais</i>	66
Σχήμα 3.15: Διάγραμμα διασποράς των μεταβλητών <i>Wt</i> και <i>RCC</i> με μικρότερο aspect ratio	67
Σχήμα 3.16: Πίνακας διαγραμμάτων διασποράς μετά τη διαγραφή του απομονωμένου σημείου	68
Σχήμα 3.17: Διάγραμμα καταλοίπων με προσαρμοσμένη <i>lowess</i> καμπύλη	68
Σχήμα 3.18: Τρισδιάστατο διάγραμμα μεταξύ των μεταβλητών <i>Ht</i> , <i>LBM</i> και <i>Wt</i>	69
Σχήμα 3.19: Διάφορες απεικονίσεις του περιστραμμένου διαγράμματος	70
Σχήμα 3.20: Η διδιάστατη απεικόνιση με την πιο ισχυρή γραμμική τάση	71
Σχήμα 3.21: Διδιάστατη απεικόνιση { <i>LBM</i> , h_{ols} }	73
Σχήμα 3.22: Διάγραμμα περίληψης και ασυσχέτιστη απεικόνιση αυτού	74
Σχήμα 3.23: Διαγράμματα αντίστροφης παλινδρόμησης με προσαρμοσμένη <i>lowess</i>	76

Κεφάλαιο 1:

Γραφικές μέθοδοι παλινδρόμησης

1.1 Εισαγωγή

Όταν κάνουμε μια ανάλυση παλινδρόμησης, σκοπός μας είναι να κατανοήσουμε τον τρόπο με τον οποίο μια εξαρτημένη μεταβλητή (response variable) εξαρτάται από μια ή περισσότερες ανεξάρτητες ή ελεγχόμενες (predictors) τυχαίες μεταβλητές. Στην περίπτωση που έχουμε μόνο μια ανεξάρτητη μεταβλητή, τότε ένα διάγραμμα δυο διαστάσεων με την ανεξάρτητη μεταβλητή στον οριζόντιο άξονα και τη μεταβλητή απόκρισης ή εξαρτημένη μεταβλητή στον κάθετο άξονα, αρκεί για να παρατηρήσουμε την εξάρτηση μεταξύ των δυο μεταβλητών. Τα πράγματα όμως γίνονται δυσκολότερα όταν έχουμε περισσότερες από μια ανεξάρτητες μεταβλητές. Παρόλα αυτά, διαγράμματα μπορούν να χρησιμοποιηθούν ώστε να απεικονίσουν την εξάρτηση μεταξύ των μεταβλητών ακόμα και σε περισσότερες από δυο διαστάσεις.

Στο σημείο αυτό, θα πρέπει να κάνουμε μια διάκριση των διαγραμμάτων σε στατικά (static) και κινητικά ή αλληλεπιδραστικά (kinetic or interactive). Τα στατικά διαγράμματα χρησιμοποιούνται όταν έχουμε μια μόνο ανεξάρτητη μεταβλητή ενώ τα κινητικά ή αλληλεπιδραστικά χρησιμεύουν για την οπτική απεικόνιση της εξάρτησης σε περίπτωση ύπαρξης περισσοτέρων μεταβλητών. Τα κινητικά διαγράμματα χρησιμοποιούν κίνηση στην οθόνη του υπολογιστή, ώστε να αποδίδουν την πληροφορία που μας ενδιαφέρει. Στα διαγράμματα αυτά μπορούμε να επεμβαίνουμε αλλάζοντας σχήμα, μέγεθος προσθέτοντας ή αφαιρώντας σημεία ή συνδέοντας διάφορα γραφήματα μεταξύ τους.

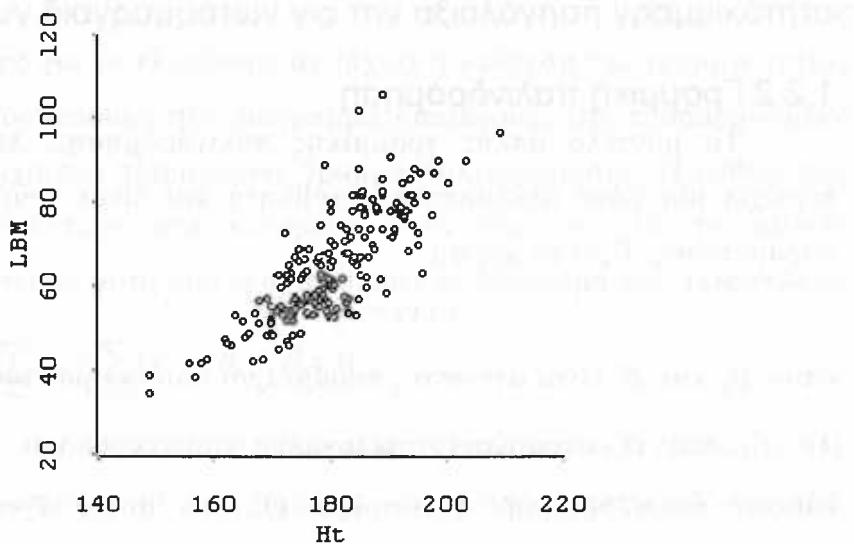
1.2 Διαγράμματα απλής παλινδρόμησης

Στην περίπτωση της απλής παλινδρόμησης έχουμε μια εξαρτημένη μεταβλητή, την y και μια ανεξάρτητη μεταβλητή που συμβολίζεται με x . Μας ενδιαφέρει η δεσμευμένη κατανομή της y δοθέντος x .

Επειδή η μελέτη της δεσμευμένης κατανομής είναι αρκετά δύσκολη η ανάλυση παλινδρόμησης συνήθως εστιάζει την προσοχή της στο χαρακτηρισμό του τρόπου με τον οποίο ο μέσος της κατανομής $y|x$, που συμβολίζεται με $E(y|x)$, και η διακύμανση της $y|x$, που συμβολίζεται με $\text{var}(y|x)$, εξαρτώνται από τη μεταβλητή x . Ο μέσος $E(y|x)$ ονομάζεται **συνάρτηση παλινδρόμησης** (regression function) ενώ η διακύμανση $\text{var}(y|x)$ καλείται **συνάρτηση διακύμανσης** (variance function).

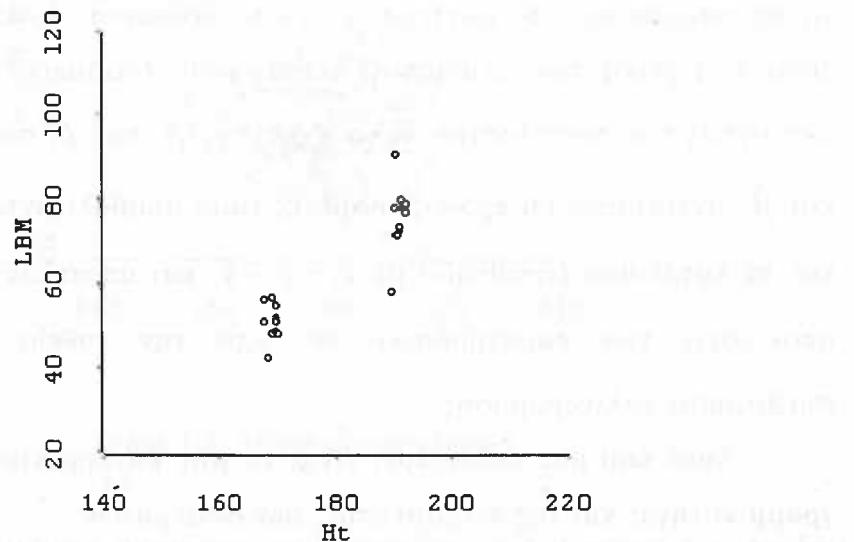
1.2.1 Διαγράμματα διασποράς

Ένας πολύ καλός τρόπος για να παρατηρήσουμε τον τρόπο με τον οποίο η κατανομή της $y|x$ μεταβάλλεται με την τιμή της x , είναι ένα διάγραμμα που στον οριζόντιο άξονα απεικονίζεται η ανεξάρτητη μεταβλητή ενώ στον οριζόντιο η εξαρτημένη. Ένα τέτοιου είδους διάγραμμα, ονομάζεται διάγραμμα διασποράς (scatterplot) και συμβολίζεται ως $\{x,y\}$. Για παράδειγμα το διάγραμμα διασποράς μεταξύ δυο μεταβλητών Ht και LBM φαίνεται στο Σχήμα 1.1. Οι μεταβλητές αυτές περιλαμβάνονται μεταξύ άλλων στο αρχείο δεδομένων «ais.lsp». Τα δεδομένα του αρχείου αυτού, αναφέρονται σε 202 αθλητές, εκ των οποίων 102 είναι άνδρες και 100 γυναίκες και έχουν συλλεχθεί από το Ινστιτούτο Αθλημάτων της Αυστραλίας (Australian Institute of Sport). Οι μεταβλητές Ht και LBM , εκφράζουν το ύψος σε εκατοστά και τη συγκέντρωση μυϊκού ιστού των αθλητών αντίστοιχα. Στη συνέχεια, θα χρησιμοποιήσουμε και τις μεταβλητές RCC και Wt , οι οποίες εκφράζουν τον αριθμό των ερυθρών αιμοσφαιρίων και το βάρος σε κιλά των αθλητών αντίστοιχα. Σημειώνουμε, ότι τόσο το πρόγραμμα Arc όσο και τα δεδομένα που χρησιμοποιούμε διατίθενται δωρεάν από το διαδίκτυο, υπό τη διεύθυνση <http://www.stat.umn.edu/arc/>.



Σχήμα 1.1: Διάγραμμα διασποράς

Έστω ότι θέλουμε να συγκρίνουμε την κατανομή της $y|(x = \tilde{x})$, όταν το \tilde{x} παίρνει δυο διαφορετικές τιμές. Για παράδειγμα, έστω ότι θέλουμε να συγκρίνουμε την κατανομή της $y|(x = 170)$ με αυτή της $y|(x = 190)$. Για να το επιτύχουμε αυτό, χρησιμοποιούμε τις τιμές της μεταβλητής y που αντιστοιχούν σε τιμές της μεταβλητής x κοντά στην τιμή \tilde{x} . Αυτή η μέθοδος ονομάζεται *τμηματοποίηση* (slicing). Το τμήμα του οριζοντίου άξονα που καταλαμβάνεται από ένα slice, ονομάζεται slice window και το εύρος αυτού καλείται εύρος παραθύρου (window width) (Σχήμα 1.2).



Σχήμα 1.2: Διάγραμμα διασποράς με slices στις τιμές $x = 170$ και $x = 190$

1.2.2 Γραμμική παλινδρόμηση

Το μοντέλο απλής γραμμικής παλινδρόμησης, λέγεται απλό γιατί περιέχει μία μόνο ανεξάρτητη μεταβλητή και είναι γραμμικό ως προς τις παραμέτρους. Έχει τη μορφή:

$$E(y) = \beta_0 + \beta_1 x, \quad (1.1)$$

όπου β_0 και β_1 είναι άγνωστες παράμετροι που εκτιμώνται από τα δεδομένα. Τα β_0 και β_1 ονομάζονται τεταγμένη (intercept) και κλίση (slope) της ευθείας παλινδρόμησης αντίστοιχα. Οι δυο αυτές άγνωστες παράμετροι εκτιμώνται από τα δεδομένα. Η τεταγμένη β_0 μας δίνει την τιμή της y όταν x παίρνει την τιμή 0 ενώ η κλίση β_1 δίνει τη μεταβολή που επέρχεται στη y όταν η x μεταβληθεί κατά μια μονάδα.

Βασική υπόθεση είναι ότι η μέση τιμή του όρου σφάλματος ισούται με το μηδέν, δηλαδή $E(\varepsilon) = 0$ και ότι η συνάρτηση διακύμανσης $\text{var}(y)$ είναι μια μη αρνητική σταθερά, δηλαδή $\text{var}(y) = \sigma^2$.

Υπό τις παραπάνω υποθέσεις, το μοντέλο της απλής γραμμικής παλινδρόμησης, μπορεί να γραφεί ως

$$y = \beta_0 + \beta_1 x + \varepsilon_i \quad (1.2)$$

για $i = 1, 2, \dots, n$, όπου n είναι το πλήθος των παρατηρήσεων.

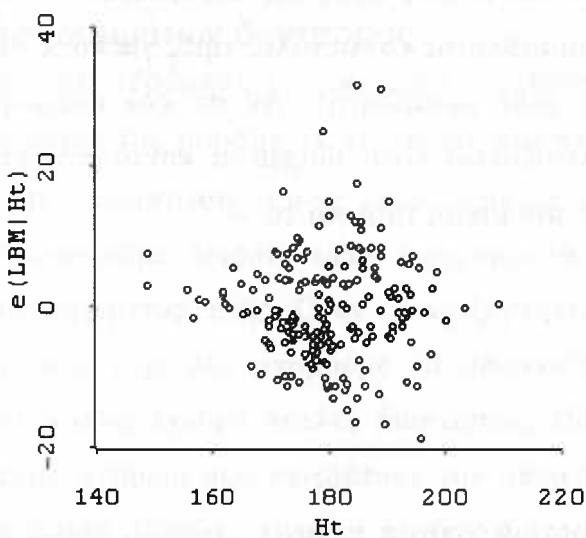
Για να εκτιμήσουμε τις άγνωστες παραμέτρους β_0 και β_1 , θα πρέπει να προσαρμόσουμε το μοντέλο (1.2) στα δεδομένα. Ένας τρόπος γι' αυτό, είναι η τεχνική των ελαχίστων τετραγώνων (ordinary least squares). Οι εκτιμήσεις των συντελεστών παλινδρόμησης β_0 και β_1 συμβολίζονται με $\hat{\beta}_0$ και $\hat{\beta}_1$, αντίστοιχα. Οι προσαρμοσμένες τιμές συμβολίζονται με $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ και τα κατάλοιπα (residuals) με $e_i = y_i - \hat{y}_i$ και αποτελούν τις κατακόρυφες αποκλίσεις των παρατηρήσεων y_i από την ευθεία της εκτιμώμενης συνάρτησης παλινδρόμησης.

Αυτό που μας ενδιαφέρει είναι να μην καταρρίπτεται η υπόθεση της γραμμικότητας και της ανεξαρτησίας των σφαλμάτων.

1.2.3 Χρήση των διαγραμμάτων για την αξιολόγηση γραμμικότητας

Ένας τρόπος για να ελέγξουμε αν ισχύει η υπόθεση της γραμμικότητας (1.1) είναι να προσθέσουμε στο διάγραμμα διασποράς, την προσαρμοσμένη με τη μέθοδο ελαχίστων τετραγώνων γραμμή παλινδρόμησης. Η ευθεία που προσαρμόζεται καλύτερα στα δεδομένα είναι σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων των κατάλοιπων $\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$.

Ένας άλλος τρόπος είναι τα διαγράμματα καταλοίπων (residual plots). Για να κατασκευάσουμε τα διαγράμματα καταλοίπων, απεικονίζουμε την ανεξάρτητη τυχαία μεταβλητή στον οριζόντιο άξονα ενώ στον κάθετο άξονα απεικονίζουμε τα κατάλοιπα. Το διάγραμμα των καταλοίπων που αντιστοιχεί στα δεδομένα του Σχήματος 1.1, παρουσιάζεται στο Σχήμα 1.3. Εάν το μοντέλο (1.2) περιγράφει ικανοποιητικά τα δεδομένα, τότε η κατανομή των σφαλμάτων ε δεν πρέπει να εξαρτάται από τη μεταβλητή x . Για να ισχύει η γραμμικότητα, θα πρέπει τα σημεία που απεικονίζονται στο διάγραμμα να είναι διεσπαρμένα γύρω από μια παράλληλη προς τον οριζόντιο άξονα ευθεία που περνά από το μηδέν χωρίς να εμφανίζουν κάποια συγκεκριμένη μορφή.



Σχήμα 1.3: Διάγραμμα καταλοίπων

Αν από τα διαγράμματα διασποράς συμπεράνουμε ότι η σχέση που συνδέει την εξαρτημένη με την ανεξάρτητη μεταβλητή δεν είναι γραμμική, τότε το

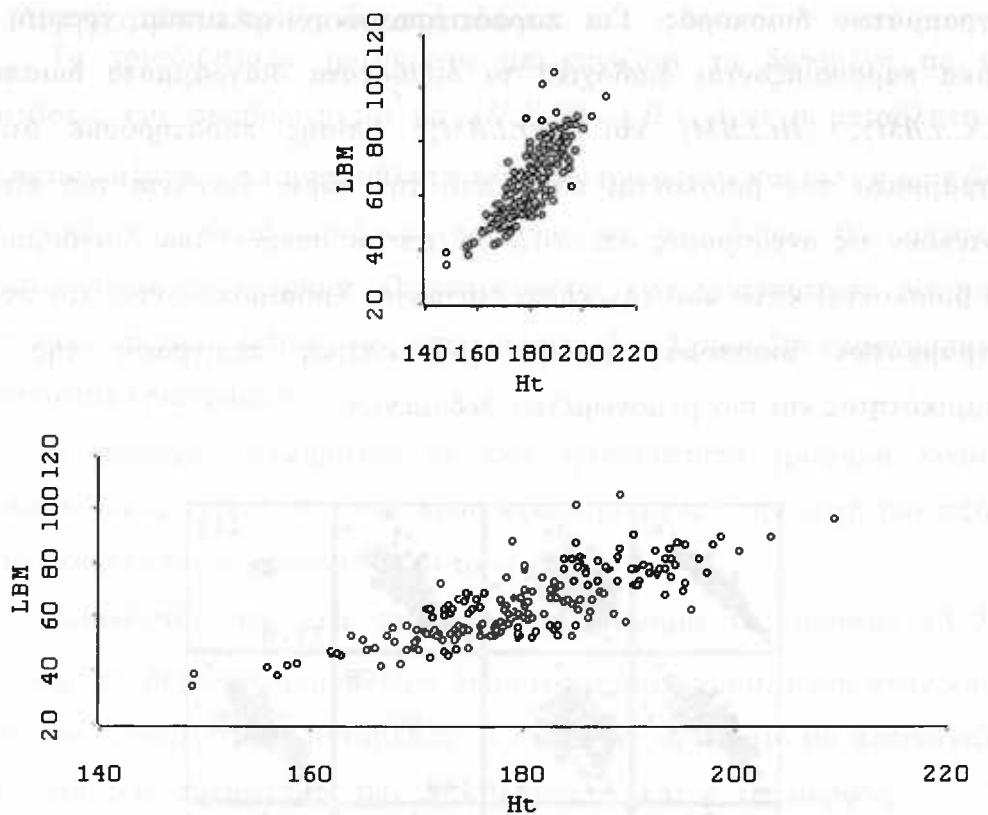
μοντέλο απλής γραμμικής παλινδρόμησης (1.2) δεν πρέπει να χρησιμοποιηθεί. Τα παραπάνω διαγράμματα δίνουν λίγη πληροφορία για τη συνάρτηση παλινδρόμησης $E(y)$. Για να ξεπεράσουμε το πρόβλημα αυτό, χρησιμοποιούμε την τεχνική του slicing που έχουμε ήδη αναφέρει.

1.3 Διαγράμματα δύο διαστάσεων

Λέγοντας *aspect ratio* εννοούμε την τιμή που προκύπτει αν διαιρέσουμε το μήκος του κάθετου άξονα του διαγράμματος με το μήκος του οριζόντιου άξονα. Είναι σημαντικό να μπορούμε να αλλάζουμε την τιμή του aspect ratio και να ανιχνεύουμε πρότυπα (patterns). Στο Σχήμα 1.4, εμφανίζεται το διάγραμμα διασποράς με διαφορετικό aspect ratio. Στο πρώτο διάγραμμα έχουμε αυξήσει τη τιμή του aspect ratio ενώ στο δεύτερο το έχουμε μειώσει. Παρατηρούμε ότι το δεύτερο διάγραμμα μπορεί να μελετηθεί καλύτερα από το πρώτο.

Για να πάρουν τα δεδομένα μας πιο επεξεργάσιμη μορφή μπορούμε να χρησιμοποιήσουμε μετασχηματισμούς δύναμης της μορφής v^* όπου η παράμετρος λ παίρνει τιμές στο διάστημα [-1,2].

Χρησιμοποιούμε μικρές τιμές της παραμέτρου λ για να εξαπλώσουμε μικρές τιμές μιας μεταβλητής και μεγάλες τιμές για το λ για να εξαπλώσουμε τις μεγάλες τιμές μιας μεταβλητής. Αν σε ένα διάγραμμα διασποράς τα σημεία που απεικονίζονται είναι μαζεμένα κοντά στο μηδέν θα πρέπει να χρησιμοποιήσουμε μια μικρή τιμή για το λ .

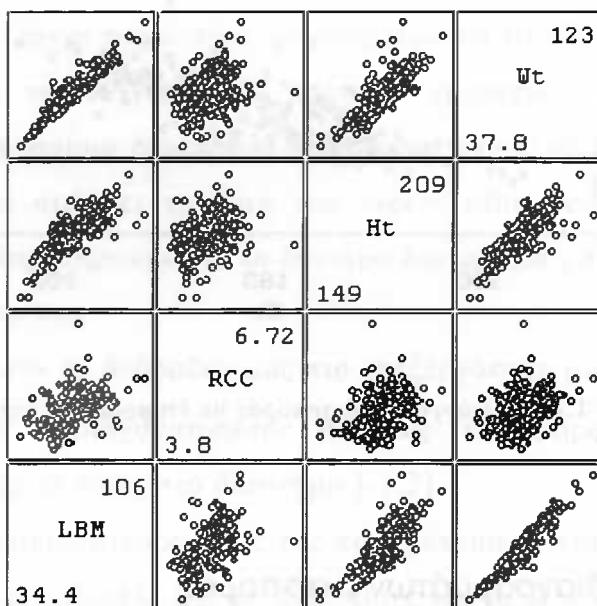


Σχήμα 1.4: Το διάγραμμα διασποράς με διαφορετικό aspect ratio

1.3.1 Πίνακες διαγραμμάτων διασποράς

Όταν έχουμε παλινδρόμηση με μια μόνο ανεξάρτητη μεταβλητή, τα διδιάστατα διαγράμματα της μορφής $\{x, y\}$ δίνουν μια πλήρη περίληψη του προβλήματος. Στην περίπτωση όμως που έχουμε δυο ανεξάρτητες μεταβλητές, χρησιμοποιούμε τρισδιάστατα διαγράμματα των ανεξάρτητων έναντι της εξαρτημένης μεταβλητής. Όταν οι ανεξάρτητες μεταβλητές είναι περισσότερες των δυο, τότε δεν μπορούμε να απεικονίσουμε γραφικά τα δεδομένα καθόλου, καθώς έχουμε πολλές διαστάσεις. Παρόλα αυτά, έχουν δημιουργηθεί κάποιες μέθοδοι που επιτρέπουν την απεικόνιση μερικών από τα δεδομένα. Μια τέτοια μέθοδος, είναι οι **πίνακες διαγραμμάτων διασποράς** (scatterplot matrices), οι οποίοι επιτρέπουν τη διατεταγμένη εμφάνιση πολλών διδιάστατων διαγραμμάτων για δεδομένα περισσοτέρων διαστάσεων. Στο Σχήμα 1.5, παρουσιάζεται ένας πίνακας διαγραμμάτων διασποράς, ο οποίος όπως παρατηρούμε, αποτελείται από μια σειρά διδιάστατων

διαγραμμάτων διασποράς. Για παράδειγμα, στην τελευταία γραμμή του πίνακα παρουσιάζονται διαδοχικά τα διδιάστατα διαγράμματα διασποράς $\{RCC,LBM\}$, $\{Ht,LBM\}$ και $\{Wt,LBM\}$. Επίσης παρατηρούμε ότι τα διαγράμματα που βρίσκονται πάνω από την κύρια διαγώνιο του πίνακα, αποτελούν τις αντίστροφες απεικονίσεις (mirror images) των διαγραμμάτων που βρίσκονται κάτω από την κύρια διαγώνιο. Χρησιμοποιώντας τον πίνακα διαγραμμάτων διασποράς έχουμε μια οπτική περιγραφή της (μη) γραμμικότητας και των μεμονωμένων δεδομένων.



Σχήμα 1.5: Πίνακας διαγραμμάτων διασποράς

1.3.2 Διαγράμματα μερικής απόκρισης

Η γραμμή του πίνακα διαγραμμάτων διασποράς, η οποία δίνει το διάγραμμα διασποράς κάθε ανεξάρτητης μεταβλητής έναντι της εξαρτημένης, αποτελεί τη σειρά με τα διαγράμματα μερικής απόκρισης (partial response plots). Τα διαγράμματα μερικής απόκρισης δείχνουν πώς εξαρτάται η εξαρτημένη μεταβλητή y από κάθε μια ανεξάρτητη μεταβλητή x , αγνοώντας τις υπόλοιπες ανεξάρτητες μεταβλητές.

1.4 Διαγράμματα τριών διαστάσεων

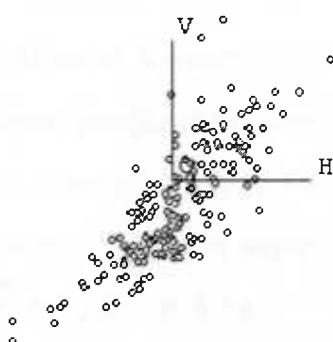
Τα τρισδιάστατα γραφήματα απεικονίζουν τα δεδομένα σε τρεις διαστάσεις και συμβολίζονται με $\{H, V, O\}$. (H) είναι η μεταβλητή του οριζόντιου άξονα, (V) η μεταβλητή του κάθετου άξονα και (O) η μεταβλητή του εκτός της σελίδας άξονα, που για να τον δούμε θα πρέπει να περιστρέψουμε το γράφημα. Περιστρέφοντας ένα τρισδιάστατο διάγραμμα παίρνουμε στατικές διδιάστατες απεικονίσεις. Στο Σχήμα 1.6 εμφανίζεται ένα τρισδιάστατο διάγραμμα.

Η περιοχή γραφήματος σε ένα τρισδιάστατο γράφημα είναι το εσωτερικό ενός κύβου, το οποίο είναι κεντραρισμένο στην αρχή των αξόνων και οι πλευρές του εκτείνονται από το -1 έως το 1.

Ξεκινώντας από ένα τρισδιάστατο γράφημα της μορφής $\{X, Y, Z\}$ μπορούμε να δημιουργήσουμε ένα διδιάστατο διάγραμμα, όπου στον κάθετο άξονα θα εμφανίζεται η μεταβλητή Y ενώ στον οριζόντιο θα απεικονίζεται ένας γραμμικός συνδυασμός των μεταβλητών X και Z της μορφής

$$\text{horizontal screen variable} = d + h = d + a(\cos \theta)X + c(\sin \theta)Z, \quad (1.3)$$

όπου τα a και c είναι οι συντελεστές κλίμακας, d είναι μια σταθερά, το θ είναι η γωνία περιστροφής του γραφήματος και h είναι ο γραμμικός συνδυασμός των μεταβλητών του οριζόντιου άξονα.

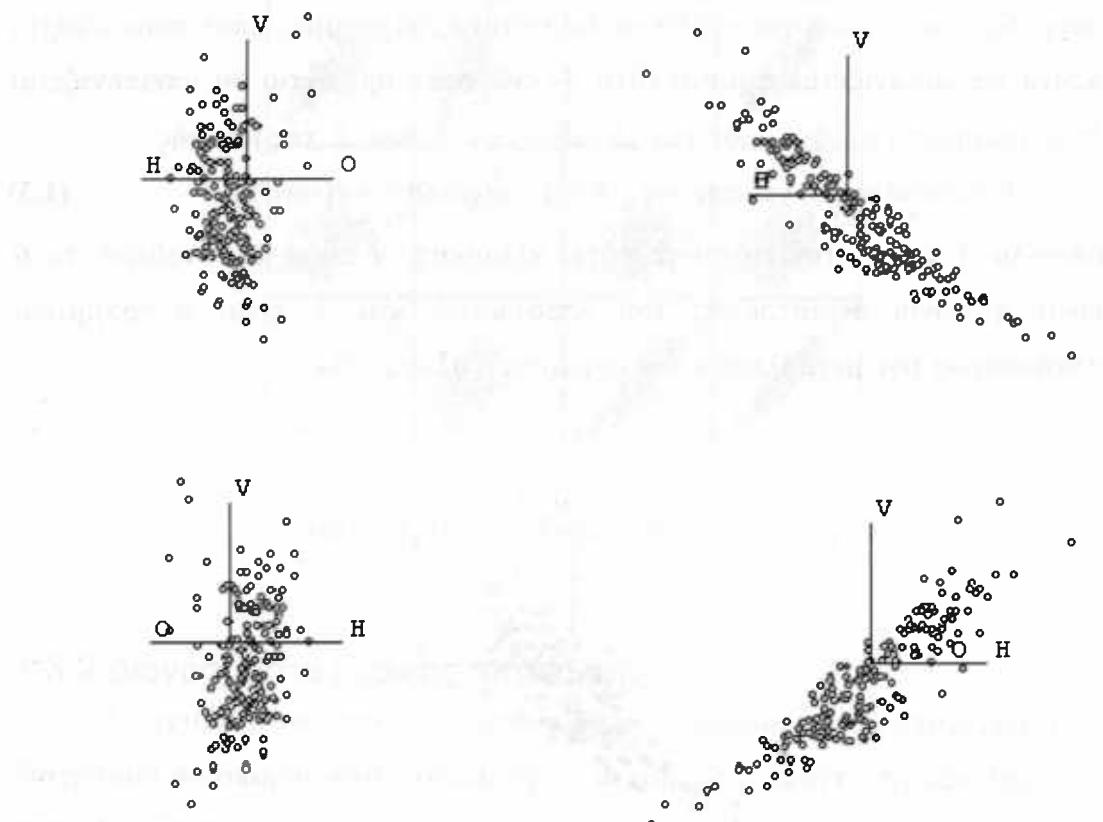


Σχήμα 1.6: Τρισδιάστατο διάγραμμα

Η στατική εικόνα ενός τρισδιάστατου γραφήματος που στον οριζόντιο άξονα έχει τη μεταβλητή Y και στον οριζόντιο τη σχέση (1.3), είναι η προβολή (projection) του τρισδιάστατου γραφήματος στο επίπεδο και

αποτελεί ένα διδιάστατο διάγραμμα. Στο Σχήμα 1.7 παρουσιάζονται τέσσερις διδιάστατες απεικονίσεις του τρισδιάστατου διαγράμματος του Σχήματος 1.6, που προέκυψαν από την περιστροφή του τρισδιάστατου διαγράμματος.

Υπάρχει περίπτωση σε ένα τρισδιάστατο γράφημα μία ισχυρή γραμμική τάση να αποκρύβει σημαντικά στοιχεία όπως οι μη γραμμικότητες. Για να απαλείψουμε τη γραμμική τάση αντικαθιστούμε τη μεταβλητή του κάθετου άξονα με την $e(V | H, O)$ που συμβολίζει τα κατάλοιπα από την παλινδρόμηση μέσω ελαχίστων τετραγώνων της μεταβλητής V με τις μεταβλητές H , O και μια σταθερά. Έτσι από ένα διάγραμμα $\{H, V, O\}$ προκύπτει ένα διάγραμμα $\{H, e(V | H, O), O\}$.



Σχήμα 1.7 Διάφορες διδιάστατες απεικονίσεις ενός τρισδιάστατου διαγράμματος

Υπάρχει περίπτωση από ένα γράφημα $\{H, V, O\}$ να χρειαστεί να πάμε σε γράφημα $\{H, V, e(O | H, V)\}$, δηλαδή να αντικαταστήσουμε τη μεταβλητή

που απεικονίζεται στον εκτός σελίδας άξονα με τα κατάλοιπα από την παλινδρόμηση μέσω ελαχίστων τετραγώνων της μεταβλητής O με την H συμπεριλαμβανομένης της σταθεράς. Αυτό θα συμβεί αν οι ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες και θα πρέπει να αντικατασταθούν με ασυσχέτιστες μεταβλητές. Οι τιμές της μεταβλητής του οριζόντιου άξονα δεν θα είναι αρκετά διεσπαρμένες. Αν η δειγματική συσχέτιση των κατάλοιπων και της μεταβλητής (H) είναι μηδέν τότε οι μεταβλητές στον οριζόντιο και τον εκτός σελίδας άξονα θα είναι ασυσχέτιστες και οι τιμές της μεταβλητής του οριζόντιου άξονα θα είναι καλά διεσπαρμένες.

1.5 Απεικόνιση γραμμικής παλινδρόμησης με δυο ανεξάρτητες μεταβλητές

Όταν έχουμε δυο ανεξάρτητες μεταβλητές, αυτό που μας ενδιαφέρει είναι πώς η κατανομή της εξαρτημένης y μεταβάλλεται σε σχέση με τις τιμές των ανεξάρτητων μεταβλητών x_1 και x_2 . Γι' αυτό χρησιμοποιούμε το περιστρεφόμενο τρισδιάστατο γράφημα $\{x_1, y, x_2\}$.

Το μοντέλο γραμμικής παλινδρόμησης με δυο ανεξάρτητες τυχαίες μεταβλητές, το οποίο είναι μια γενίκευση του μοντέλου (1.2), δίνεται από τη σχέση

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \quad (1.4)$$

Τα β_0 , β_1 και β_2 είναι οι άγνωστοι συντελεστές παλινδρόμησης ενώ τα σφάλματα ε θεωρούνται ανεξάρτητα μεταξύ τους και ανεξάρτητα των μεταβλητών x_1 , x_2 , έχουν μέση τιμή μηδέν και σταθερή διακύμανση σ^2 . Το μοντέλο (1.4) μπορεί να γραφεί και με τη μορφή

$$y | \mathbf{x} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \varepsilon, \quad (1.5)$$

όπου $\mathbf{x} = (x_1, x_2)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$.

Η κατανομή της $y | \mathbf{x}$ εξαρτάται από το διάνυσμα μεταβλητών \mathbf{x} μόνο μέσα από το γραμμικό συνδυασμό $\boldsymbol{\beta}^T \mathbf{x}$, δηλαδή η κατανομή της $y | \mathbf{x}$ είναι ίδια με την κατανομή της $y | \boldsymbol{\beta}^T \mathbf{x}$ για όλες τις τιμές του \mathbf{x} και $E(y | \mathbf{x}) = E(y | \boldsymbol{\beta}^T \mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}$.

1.5.1 Το ιδανικό διάγραμμα περίληψης

Υποθέτουμε ότι γνωρίζουμε το γινόμενο $c\beta$ για κάποια μη μηδενική τιμή c . Η υπόθεση αυτή γίνεται γιατί γραφικά μπορούμε να εκτιμήσουμε την ποσότητα $c\beta$ ενώ το β μόνο του όχι. Στη συνέχεια υπολογίζουμε μια νέα ανεξάρτητη μεταβλητή $h^* = c\beta^T \mathbf{x}$ για κάθε τιμή του διανύσματος \mathbf{x} και ξαναγράφουμε τη συνάρτηση παλινδρόμησης ως

$$E(y | h^*) = \beta_0 + c^{-1} h^*. \quad (1.6)$$

Συνεπώς η γνώση της ποσότητας $c\beta$ μας επιτρέπει να μετατρέψουμε το αρχικό μοντέλο παλινδρόμησης σε ένα μοντέλο απλής γραμμικής παλινδρόμησης χωρίς να χάσουμε καμία πληροφορία.

Τα δεδομένα που χρησιμοποιούνται για να προσαρμόσουμε το μοντέλο (1.6), μπορούμε να τα απεικονίσουμε γραφικά σε ένα διδιάστατο γράφημα της μορφής $\{h^*, y\}$, το οποίο ονομάζεται *ιδανικό διάγραμμα περίληψης* (ideal summary plot).

Επειδή είναι σχεδόν απίθανο να γνωρίζουμε την τιμή της ποσότητας $c\beta$ δεν είναι εύκολο να κατασκευάσουμε το ιδανικό διάγραμμα περίληψης. Μπορούμε όμως να προσδιορίσουμε μια εκτίμηση h του h^* μέσω ενός τρισδιάστατου γραφήματος.

1.5.2 Προσαρμογή της ευθείας παλινδρόμησης

Έστω ότι έχουμε ένα τρισδιάστατο γράφημα $\{x_1, E(y | x), x_2\}$, το οποίο περιστρέφουμε μέχρις ότου στην οθόνη να εμφανίζεται ένα διδιάστατο γράφημα στο οποίο τα σημεία που απεικονίζονται να αποτελούν μια απλή ευθεία γραμμή. Αυτό σημαίνει ότι έχουμε προσαρμόσει μια ευθεία παλινδρόμησης με το μάτι. Η προσαρμογή αυτή είναι εύκολη όταν δεν υπάρχουν σφάλματα, δηλαδή όταν $\sigma^2 = 0$.

Στην περίπτωση που ισχύει $\sigma^2 > 0$ κατασκευάζουμε το γράφημα $\{x_1, y, x_2\}$ αντί του $\{x_1, E(y | x), x_2\}$. Τα σημεία που απεικονίζονται σε αυτό δεν πέφτουν ακριβώς στο επίπεδο εξαιτίας της ύπαρξης των σφαλμάτων. Εάν περιστρέψουμε το τρισδιάστατο αυτό γράφημα, παρατηρούμε ότι από ορισμένες πλευρές ανιχνεύουμε ισχυρή γραμμικότητα ενώ από άλλες ασθενή.

Εμάς μας ενδιαφέρει να παρατηρήσουμε την όσο το δυνατόν ισχυρότερη γραμμική σχέση.

Μια τρίτη εναλλακτική περίπτωση είναι να προσαρμόσουμε την ευθεία παλινδρόμησης, ελαχιστοποιώντας το άθροισμα των τετραγώνων των διαφορών μεταξύ των παρατηρούμενων y και των αντίστοιχων προσαρμοσμένων τιμών. Χρησιμοποιούμε δηλαδή τη μέθοδο των ελαχίστων τετραγώνων. Το επίπεδο ελαχίστων τετραγώνων, περιλαμβάνει όλα τα σημεία $\{x_1, \hat{y}, x_2\}$, όπου $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ είναι οι προσαρμοσμένες τιμές και $\hat{\beta}_0$, $\hat{\beta}_1$ και $\hat{\beta}_2$ είναι οι εκτιμημένες τιμές, μέσω της μεθόδου ελαχίστων τετραγώνων, των άγνωστων παραμέτρων. Συνεπώς προσδιορίζουμε ένα γραμμικό συνδυασμό $h_{ols} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Αν το $\hat{\beta}$ είναι καλή εκτίμηση του β , τότε η εξίσωση (1.6) θα ικανοποιείται, έχοντας αντικαταστήσει το h^* με το h_{ols} .

Περιστρέφοντας ένα τρισδιάστατο γράφημα γύρω από τον κάθετο άξονα, παίρνουμε διδιάστατα γραφήματα της μορφής $\{h, y\}$, όπου $h = b_1 x_1 + b_2 x_2 = \mathbf{b}' \mathbf{x}$ για κάποιο διάνυσμα \mathbf{b} . Σταματάμε την περιστροφή εκεί που εμφανίζεται η πιο ισχυρή γραμμική τάση. Αν βρούμε $\mathbf{b} \approx c\beta$ τότε η σχέση (1.6) θα ικανοποιείται αντικαθιστώντας το h^* με το h .

Αν θεωρήσουμε ότι το h αποτελεί μια καλή προσέγγιση του $h^* = c\beta' \mathbf{x}$ τότε το γράφημα περίληψης $\{h, y\}$ είναι επαρκές και μας δείχνει με ποιο τρόπο η κατανομή της $y | \mathbf{x}$ μεταβάλλεται με την τιμή του διανύσματος \mathbf{x} .

Για να ελέγξουμε αν η κατανομή της $y | \mathbf{x}$ είναι ανεξάρτητη του \mathbf{x} χρησιμοποιούμε μια απλή διαδικασία που στηρίζεται σε μια απλή διδιάστατη απεικόνιση. Ο οριζόντιος άξονας αποτελεί το γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που είναι ασυσχέτιστες με το h . Στη συνέχεια, κατασκευάζουμε το διάγραμμα $\{h_u, y\}$ που ονομάζεται *ασυσχέτιστη διδιάστατη απεικόνιση* (uncorrelated 2D view), όπου h_u είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών. Εάν στην απεικόνιση αυτή δεν παρατηρείται κάποιο πρότυπο, τότε το διάγραμμα $\{h, y\}$ θεωρείται επαρκές γράφημα περίληψης των δεδομένων.

1.5.3 Κατανομή των ανεξάρτητων μεταβλητών

Πολλές φορές, ακόμη και αν η συνάρτηση παλινδρόμησης $E(y | \mathbf{x})$ είναι γραμμική συνάρτηση του διανύσματος \mathbf{x} , σε κάποιες από τις διδιάστατες απεικονίσεις ενός τρισδιάστατου γραφήματος μπορεί να εμφανίζονται καθαρές μη γραμμικές σχέσεις. Αν οι συναρτήσεις $E(x_1 | x_2)$ και $E(x_2 | x_1)$, δηλαδή η συνάρτηση παλινδρόμησης της x_1 με τη x_2 και η συνάρτηση παλινδρόμησης της x_2 με τη x_1 , είναι και οι δυο γραμμικές, τότε όλες οι στατικές διδιάστατες απεικονίσεις του γραφήματος $\{x_1, y, x_2\}$ θα παρουσιάζουν γραμμική σχέση. Αν μια από αυτές είναι μη γραμμική, τότε κάποιες απεικονίσεις μπορεί να παρουσιάζουν μη γραμμική σχέση, ακόμη και αν ισχύει το γραμμικό μοντέλο παλινδρόμησης (1.4).

Το γράφημα $\{x_1, y\}$ χρησιμοποιείται για να εκτιμήσει τη συνάρτηση μερικής παλινδρόμησης (partial regression function) $E(y | x_1)$. Υπάρχουν περιπτώσεις όπου ενώ η $E(y | \mathbf{x}) = E(y | x_1, x_2)$ είναι γραμμική συνάρτηση του διανύσματος \mathbf{x} , η $E(y | x_1)$ είναι μη γραμμική συνάρτηση της μεταβλητής x_1 .

Συχνά παρατηρείται το φαινόμενο σε ένα διδιάστατο διάγραμμα να εμφανίζεται μη σταθερή διακύμανση. Αυτό σημαίνει ότι η συνάρτηση μερικής διακύμανσης (partial variance function) $\text{var}(y | x_2)$ εξαρτάται από τη μεταβλητή x_2 , επειδή υπάρχει γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών. Η συμπεριφορά του διαγράμματος μερικής περίληψης $\{x_1, y\}$ εξαρτάται από τη συνάρτηση παλινδρόμησης $E(x_2 | x_1)$ ακόμη και αν ισχύει το γραμμικό μοντέλο (1.4). Η συνάρτηση παλινδρόμησης $E(y | \mathbf{x})$ θα είναι γραμμική αν η $E(x_2 | x_1)$ είναι γραμμική ως προς το x_1 και ισχύει το γραμμικό μοντέλο. Παρόμοια, η συνάρτηση παλινδρόμησης για το διάγραμμα $\{x_2, y\}$ εξαρτάται από την $E(x_1 | x_2)$ που είναι η συνάρτηση παλινδρόμησης για το διάγραμμα $\{x_2, x_1\}$. Στην περίπτωση που η σχέση μεταξύ των ανεξάρτητων μεταβλητών δεν είναι μονότονη, είναι πιθανό η $E(x_2 | x_1)$ να είναι γραμμική ενώ η $E(x_1 | x_2)$ να είναι μη γραμμική.

Τα συμπεράσματα που ισχύουν όταν έχουμε δυο ανεξάρτητες μεταβλητές γενικεύονται και στην περίπτωση που έχουμε περισσότερες των δυο ανεξάρτητων μεταβλητών.

Κεφάλαιο 2:

Έλεγχος της δομικής διάστασης της παλινδρόμησης

2.1 Εισαγωγή

Στην ενότητα αυτή θα αναφερθούμε στον τρόπο με τον οποίο μπορούμε να απεικονίσουμε και να συνοψίσουμε δεδομένα χωρίς να υποθέσουμε την ισχύ ενός συγκεκριμένου μοντέλου.

2.1.1 Γενικά τρισδιάστατα διαγράμματα απόκρισης

Υποθέτουμε ότι έχουμε ένα πρόβλημα πολλαπλής παλινδρόμησης με εξαρτημένη μεταβλητή την y και p ανεξάρτητες ή ερμηνευτικές μεταβλητές που αποτελούν το διάνυσμα $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$. Ο μικρότερος αριθμός των γραμμικών συνδυασμών του διανύσματος \mathbf{x} που χρειάζονται για να χαρακτηριστεί η παλινδρόμηση χωρίς απώλεια πληροφορίας, ονομάζεται *δομική διάσταση* (structural dimension) της παλινδρόμησης (Cook and Weisberg, 1999). Η δομική διάσταση είναι πάντοτε ένας ακέραιος αριθμός με τιμές μεταξύ του 0 και του p . Στην περίπτωση που έχουμε παλινδρόμηση με δυο ανεξάρτητες μεταβλητές, τότε η δομική διάστασή της θα είναι 0, 1, ή 2 και θα λέμε ότι το τρισδιάστατο διάγραμμα έχει 0D δομή, 1D δομή ή 2D δομή.

i) 0D δομή

Είναι γνωστό ότι αν η κατανομή της $y | \mathbf{x}$ δεν εξαρτάται από την τιμή του διανύσματος \mathbf{x} , τότε και η συνάρτηση παλινδρόμησης $E(y | \mathbf{x})$ και η συνάρτηση διακύμανσης $\text{var}(y | \mathbf{x})$ δεν θα επηρεάζονται από τις τιμές του διανύσματος \mathbf{x} . Αυτό σημαίνει ότι αν περιστρέψουμε το γράφημα γύρω από

τον κάθετο άξονα, δεν παρατηρούμε κανένα συστηματικό πρότυπο σε καμιά από τις διδιάστατες απεικονίσεις. Επίσης, στην περίπτωση της τιμηματοποίησης (slicing), τα σημεία σε οποιοδήποτε slice οποιασδήποτε διδιάστατης απεικόνισης θα κατανέμονται με τον ίδιο τρόπο. Στην περίπτωση αυτή, θα λέμε ότι το τρισδιάστατο διάγραμμα παρουσιάζει δομή μηδενικής διάστασης (0-dimensional (0D) structure). Αν η $y | \mathbf{x}$ δεν εξαρτάται από το \mathbf{x} , τότε έχουμε πρότυπο 0D, αφού κανένας γραμμικός συνδυασμός του \mathbf{x} δεν παρέχει πληροφορία για το y . Για την παρουσίαση των δεδομένων αρκεί να χρησιμοποιήσουμε ένα ιστόγραμμα της μεταβλητής απόκρισης (y). Αν σε κάποια διδιάστατη απεικόνιση ενός τρισδιάστατου διαγράμματος εμφανίζεται κάποιο συστηματικό πρότυπο απορρίπτουμε τη μηδενική δομική διάσταση.

Αν το πρότυπο αυτό είναι καμπυλοειδές τότε η συνάρτηση παλινδρόμησης είναι μη γραμμική ως προς το διάνυσμα \mathbf{x} . Αν έχει σχήμα βεντάλιας (fan-shaped) τότε η συνάρτηση διακύμανσης μεταβάλλεται με το \mathbf{x} .

ii) 1D δομή

Αν απαιτείται ένας μόνο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών για να συνοψίσουμε την εξάρτηση της $y | \mathbf{x}$ από το x τότε το γράφημα $\{x_1, y, x_2\}$ και το αντίστοιχο πρόβλημα παλινδρόμησης, λέμε ότι έχουν δομή μιας διάστασης (1-dimensional (1D) structure).

Το πιο γενικό μοντέλο με 1D δομή είναι το

$$y | \mathbf{x} = f(\boldsymbol{\beta}^T \mathbf{x}) + \sigma(\boldsymbol{\beta}^T \mathbf{x})\varepsilon. \quad (2.1)$$

Η συνάρτηση παλινδρόμησης και η συνάρτηση διακύμανσης όπως φαίνεται από τη σχέση (2.1) εξαρτώνται από τον ίδιο γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών $\boldsymbol{\beta}^T \mathbf{x}$.

Ειδικές περιπτώσεις του μοντέλου (2.1) είναι οι (2.2) και (2.3)

$$y | \mathbf{x} = f(\boldsymbol{\beta}^T \mathbf{x}) + \sigma\varepsilon \quad (2.2)$$

$$y | \mathbf{x} = \sigma(\boldsymbol{\beta}^T \mathbf{x})\varepsilon. \quad (2.3)$$

Για το μοντέλο (2.2) ισχύει $E(y | \mathbf{x}) = f(\boldsymbol{\beta}^T \mathbf{x})$ και $\text{var}(y | \mathbf{x}) = \sigma^2 \text{var}(\varepsilon) = \sigma^2$ όπου f είναι μια γνωστή ή άγνωστη συνάρτηση, σ είναι η συνήθως άγνωστη και κοινή για όλα τα σφάλματα τυπική απόκλιση και ε είναι μια τυχαία

μεταβλητή με μέση τιμή 0 και διακύμανση 1. Η συνάρτηση f ονομάζεται μέση συνάρτηση πυρήνα (kernel mean function). Το διδιάστατο γράφημα $\{\beta^T x, y\}$ είναι ένα ιδανικό διάγραμμα περίληψης γι' αυτό το μοντέλο, γιατί περιέχει όλη την πληροφορία που το διάνυσμα x διαθέτει για τη y . Αν η f είναι γραμμική δηλαδή $f(\beta^T x) = \beta_0 + c\beta^T x$, τότε το γραμμικό μοντέλο (1.5) αποτελεί ειδική περίπτωση του (2.2).

Για το μοντέλο (2.3) η $\sigma(\beta^T x)$ είναι μια μη αρνητική συνάρτηση η οποία μπορεί να έχει διαφορετική τιμή για κάθε τιμή του $\beta^T x$, έτσι ώστε κάθε παρατήρηση να έχει τη δική της τυπική απόκλιση σφάλματος. Η συνάρτηση $\sigma(\beta^T x)$ καλείται συνάρτηση διακύμανσης πυρήνα (kernel variance function). Στη σχέση (2.3), η συνάρτηση παλινδρόμησης είναι σταθερή και ίση με 0 για όλες τις τιμές του διανύσματος x ενώ η συνάρτηση διακύμανσης μεταβάλλεται με το x .

Στην περίπτωση που το πρόβλημα παλινδρόμησης που μελετάμε έχει 1D δομή, το διδιάστατο γράφημα $\{\beta^T x, y\}$ είναι ένα ιδανικό διάγραμμα περίληψης γι' αυτό το μοντέλο. Για να αποφασίσουμε αν η δομή 1D είναι κατάλληλη για ένα τρισδιάστατο γράφημα, το περιστρέφουμε ώστε να βλέπουμε τη διδιάστατη απεικόνιση με το πιο εμφανές πρότυπο, το οποίο μπορεί να είναι γραμμικό, μη γραμμικό ή να δείχνει μεταβαλλόμενη διακύμανση. Υποθέτουμε ότι $\beta^T x$ είναι η horizontal screen variable. Αν η 1D δομή είναι κατάλληλη και ισχύει $b^T x \approx c\beta^T x$ για κάποια μη μηδενική σταθερά c , τότε το διάγραμμα $\{\beta^T x, y\}$ θα περιέχει τη μεγαλύτερη πληροφορία για την y . Αυτό μπορεί να ελεγχθεί χωρίζοντας το γράφημα σε slices και παρατηρώντας τις αντίστοιχες ασυσχέτιστες διδιάστατες απεικονίσεις όπως έχουμε περιγράψει στο προηγούμενο κεφάλαιο. Τα slices θα πρέπει να δείχνουν μια οριζόντια διασπορά των σημείων. Σε αντίθετη περίπτωση, είτε το $c\beta^T x$ δεν είναι ανάλογο του $\beta^T x$ είτε το τρισδιάστατο γράφημα δεν έχει 1D δομή.

iii) 2D δομή

Αν οι ασυσχέτιστες διδιάστατες απεικονίσεις δείχνουν εξάρτηση, ανεξάρτητα από την επιλογή του \mathbf{b} , τότε λέμε ότι το γράφημα παρουσιάζει δομή δυο διαστάσεων (2-dimensional (2D) structure). Για να κατανοήσουμε την εξάρτηση της $y | \mathbf{x}$ από το διάνυσμα \mathbf{x} πρέπει να γνωρίζουμε την τιμή δυο ανεξάρτητων γραμμικών συνδυασμών, έστω $\beta^T \mathbf{x}$ και $\mathbf{a}^T \mathbf{x}$. Ο γραμμικός συνδυασμός $\beta^T \mathbf{x}$ απαιτείται για τον ορισμό της συνάρτησης παλινδρόμησης και ο γραμμικός συνδυασμός $\mathbf{a}^T \mathbf{x}$ απαιτείται για τον ορισμό της συνάρτηση διακύμανσης. Οπότε το μοντέλο με 2D δομή θα έχει τη μορφή

$$y | \mathbf{x} = f(\beta^T \mathbf{x}) + \sigma(\mathbf{a}^T \mathbf{x})\varepsilon.$$

Ένα άλλο μοντέλο με 2D δομή είναι το

$$y | \mathbf{x} = f(\beta^T \mathbf{x}, \mathbf{a}^T \mathbf{x}) + \sigma\varepsilon, \quad (2.4)$$

όπου η συνάρτηση παλινδρόμησης είναι $E(y | \mathbf{x}) = f(\beta^T \mathbf{x}, \mathbf{a}^T \mathbf{x})$ και εξαρτάται από δυο γραμμικούς μετασχηματισμούς ενώ η συνάρτηση διακύμανσης είναι σταθερή, $\text{var}(y | \mathbf{x}) = \sigma^2$.

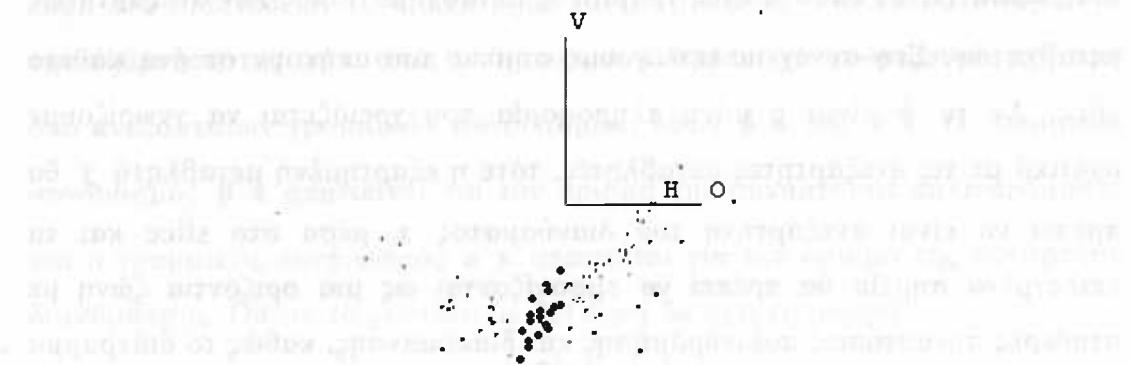
2.1.2 Έλεγχος ενός εκτιμημένου διαγράμματος περίληψης

Σε ένα τρισδιάστατο διάγραμμα, συνήθως είναι εύκολο να παρατηρήσουμε αν αυτό έχει μηδενική ή μεγαλύτερη δομική διάσταση. Αν κάποια διδιάστατη απεικόνιση έχει είτε μη σταθερή συνάρτηση παλινδρόμησης είτε μη σταθερή συνάρτηση διακύμανσης, τότε θα πρέπει να εγκαταλείψουμε την υπόθεση της μηδενικής δομικής διάστασης. Τα πράγματα δυσκολεύονται όταν θέλουμε να επιλέξουμε ανάμεσα σε δομή 1D και 2D. Έστω ότι έχουμε ένα διάγραμμα $\{y, \mathbf{b}^T \mathbf{x}\}$, το οποίο αποτελεί μια εκτίμηση του επαρκούς διαγράμματος περίληψης. Γνωρίζουμε ότι αν ισχύει η 1D δομή τότε η εξαρτημένη μεταβλητή y είναι ανεξάρτητη του διανύσματος \mathbf{x} δοθέντος ενός γραμμικού μετασχηματισμού $h^* = \beta^T \mathbf{x}$. Το γεγονός αυτό αποτελεί τη βάση για να ελέγξουμε εάν ένα διάγραμμα περίληψης χάνει σημαντική πληροφορία ή όχι. Στην ουσία ελέγχουμε το διάγραμμα για να δούμε εάν υπάρχει πληροφορία που να διαψεύδει την ανεξαρτησία της y από το \mathbf{x} δοθέντος του γραμμικού συνδυασμού $\mathbf{b}^T \mathbf{x}$.

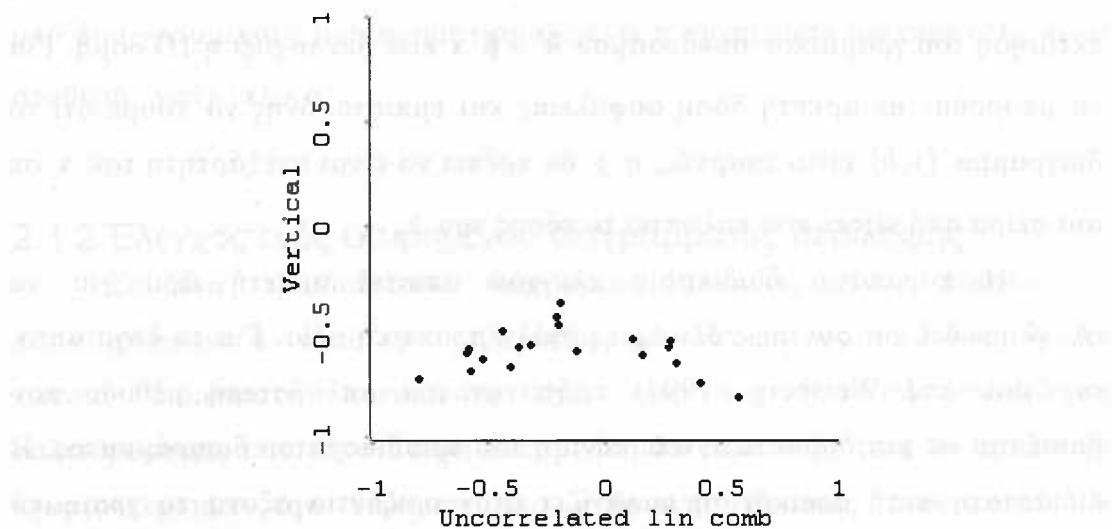
Για να αποφασίσουμε σχετικά με τη δομική διάσταση ενός προβλήματος παλινδρόμησης ή ενός τρισδιάστατου διαγράμματος ακολουθούμε την παρακάτω διαδικασία. Υποθέτουμε ότι έχουμε ένα διάγραμμα $\{y, h\}$, όπου h ένας γραμμικός μετασχηματισμός των ανεξάρτητων μεταβλητών. Στη συνέχεια επιλέγουμε σημεία που ανήκουν σε ένα κάθετο slice. Αν το h είναι η μόνη πληροφορία που χρειάζεται να γνωρίζουμε σχετικά με τις ανεξάρτητες μεταβλητές, τότε η εξαρτημένη μεταβλητή y θα πρέπει να είναι ανεξάρτητη του διανύσματος x μέσα στο slice και τα επιλεγμένα σημεία θα πρέπει να εμφανίζονται ως μια οριζόντια ζώνη με σταθερές συναρτήσεις παλινδρόμησης και διακύμανσης, καθώς το διάγραμμα περιστρέφεται. Σε περίπτωση που αντιληφθούμε σε κάποιο slice ότι υπάρχει κάποιου είδους εξάρτηση της y από το x , τότε το εκτιμημένο διάγραμμα περίληψης δεν περιλαμβάνει όλη την πληροφορία σχετικά με την κατανομή της $y|x$. Αυτό σημαίνει ότι είτε ο γραμμικός συνδυασμός h είναι κακή εκτίμηση του γραμμικού συνδυασμού $h^* = \beta^T x$ είτε δεν ισχύει η 1D δομή. Για να μπορούμε με αρκετή δόση ασφάλειας και εμπιστοσύνης να πούμε ότι το διάγραμμα $\{y, h\}$ είναι επαρκές, η y θα πρέπει να είναι ανεξάρτητη του x σε μια σειρά από slices που καλύπτει το εύρος του h .

Η παραπάνω διαδικασία ελέγχου απαιτεί αρκετή ώρα για να ολοκληρωθεί και συνεπώς δεν έχει μεγάλη πρακτική αξία. Για το λόγο αυτό, οι Cook and Weisberg (1994) πρότειναν μια απλούστερη μέθοδο που βασίζεται σε μια διδιάστατη απεικόνιση του τρισδιάστατου διαγράμματος. Η διδιάστατη αυτή απεικόνιση εμφανίζει στον οριζόντιο άξονα το γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που είναι ασυσχέτιστος με το γραμμικό συνδυασμό h . Ο γραμμικός αυτός συνδυασμός συμβολίζεται με h_{unc} ενώ το διάγραμμα $\{y, h_{unc}\}$ καλείται ασυσχέτιστη διδιάστατη απεικόνιση (uncorrelated 2D view). Τα σημεία που ανήκουν στο slice που έχουμε επιλέξει, εμφανίζονται στην ασυσχέτιστη διδιάστατη απεικόνιση. Αν στην ασυσχέτιστη διδιάστατη απεικόνιση δεν εμφανίζεται καμιά εντός των slices μορφή, τότε το διδιάστατο διάγραμμα $\{y, h\}$ αποτελεί καλό διάγραμμα περίληψης και υποστηρίζει την 1D δομή. Στο Σχήμα 2.1 παρουσιάζεται το διάγραμμα περίληψης με τονισμένα τα σημεία που ανήκουν σε ένα slice ενώ

στο Σχήμα 2.2 δίνεται η ασυσχέτιστη απεικόνιση, η οποία παρουσιάζει μόνο τα τονισμένα σημεία του Σχήματος 2.1.



Σχήμα 2.1: Διάγραμμα περίληψης



Σχήμα 2.2: Ασυσχέτιστη απεικόνιση

2.1.3 Δομική διάσταση παλινδρόμησης με πολλές ανεξάρτητες μεταβλητές

Είδαμε στην προηγούμενη ενότητα, ότι στην περίπτωση που έχουμε δυο ανεξάρτητες μεταβλητές, το διάγραμμα περίληψης μπορεί να είναι είτε ένα ιστόγραμμα, όταν η παλινδρόμηση έχει μηδενική δομική διάσταση, είτε ένα διδιάστατο διάγραμμα διασποράς της εξαρτημένης μεταβλητής έναντι του

γραμμικού συνδυασμού των ανεξάρτητων μεταβλητών, όταν έχουμε 1D δομή, είτε ένα τρισδιάστατο διάγραμμα όταν έχουμε 2D δομή. Στην περίπτωση που έχουμε περισσότερες των δυο ανεξάρτητες μεταβλητές, δηλαδή ισχύει $p \geq 2$, τότε η δομική διάσταση του προβλήματος μπορεί να είναι οποιοσδήποτε ακέραιος αριθμός μεταξύ του 0 και του p . Και σε αυτή την περίπτωση, τα πράγματα είναι παρόμοια. Δηλαδή αν έχουμε k τάξης δομική διάσταση, τότε η εξαρτημένη μεταβλητή εξαρτάται από το διάνυσμα των ανεξάρτητων μεταβλητών μέσω k γραμμικών συνδυασμών αυτών, οπότε για να δούμε πλήρως την εξάρτηση θα πρέπει να κατασκευάσουμε ένα $(k+1)$ -διάστατο διάγραμμα διασποράς. Επειδή δεν μπορούμε να κατασκευάζουμε διαγράμματα περισσοτέρων των τριών διαστάσεων, θα πρέπει να βασιζόμαστε μόνο σε διδιάστατα και τρισδιάστατα διαγράμματα. Παρόλα αυτά, για να είναι χρήσιμα τα διδιάστατα και τρισδιάστατα διαγράμματα θα πρέπει να μην διαστρεβλώνουν τις μεγαλύτερων διαστάσεων σχέσεις. Άλλωστε, αν και η πολυπλοκότητα του προβλήματος παλινδρόμησης μπορεί να αυξηθεί με τον αριθμό των ανεξάρτητων μεταβλητών, γενικά στα περισσότερα πρακτικά προβλήματα αρκούν οι δυο διαστάσεις.

Έστω τώρα ότι το πρόβλημα παλινδρόμησης έχει 1D δομή, οπότε η ανεξάρτητη μεταβλητή y εξαρτάται από το διάνυσμα x μόνο μέσω του γραμμικού συνδυασμού $\beta^T x$. Αν οι ανεξάρτητες μεταβλητές σχετίζονται μεταξύ τους γραμμικά, τότε η συνάρτηση παλινδρόμησης για κάθε ανεξάρτητη μεταβλητή x_j , $j = 1, 2, \dots, p$, $E(x_j | \beta^T x)$, θα πρέπει να είναι γραμμική συνάρτηση του γραμμικού μετασχηματισμού $\beta^T x$, δηλαδή

$$E(x_j | \beta^T x) = a_j + b_j \beta^T x.$$

Υποθέτουμε τώρα ότι έχουμε ένα πρόβλημα παλινδρόμησης, στο οποίο η y εξαρτάται από το x μέσω δυο γραμμικών συνδυασμών $\beta^T x$ και $a^T x$. Και πάλι η συνάρτηση παλινδρόμησης $E(x_j | \beta^T x, a^T x)$ θα πρέπει να είναι γραμμική συνάρτηση των $\beta^T x$ και $a^T x$, δηλαδή να ισχύει

$$E(x_j | \beta^T x, a^T x) = a_j + b_j \beta^T x + c_j a^T x,$$

για όλες τις ανεξάρτητες μεταβλητές x_j , $j = 1, 2, \dots, p$.

Αν η παλινδρόμηση έχει k D δομή, τότε απαιτούμε η συνάρτηση παλινδρόμησης $E(x_i | \beta_1^T x, \dots, \beta_k^T x)$ να είναι επίσης γραμμική συνάρτηση των ανεξάρτητων μεταβλητών. Οι τρεις παραπάνω συνθήκες, καλούνται γραμμικώς συσχετισμένες ανεξάρτητες μεταβλητές (linearly related predictors).

Η συνθήκη των γραμμικώς συσχετισμένων ανεξάρτητων μεταβλητών δεν μπορεί να ελεγχθεί ευθέως καθώς οι συντελεστές β που εμπλέκονται στη συνάρτηση παλινδρόμησης είναι άγνωστοι. Παρόλα αυτά υπάρχουν κάποιοι έμμεσοι τρόποι για τον έλεγχο της υπόθεσης αυτής. Η συνθήκη θα ισχύει αν προσεγγιστικά κάθε διάγραμμα σε έναν πίνακα διαγραμμάτων διασποράς έχει συνάρτηση παλινδρόμησης, η οποία είναι είτε γραμμική είτε στη χειρότερη περίπτωση δεν εμφανίζει ιδιαίτερη καμπυλότητα. Αν κάποιο διάγραμμα παρουσιάζει καμπυλοειδή συνάρτηση παλινδρόμησης, τότε η συνθήκη μάλλον δεν θα ισχύει. Στην περίπτωση αυτή, οι Cook and Weisberg (1999) προτείνουν το μετασχηματισμό των ανεξάρτητων μεταβλητών έτσι ώστε αυτές να ακολουθούν την πολυμεταβλητή κανονική κατανομή. Όταν οι μεταβλητές ακολουθούν την πολυμεταβλητή κανονική κατανομή, τότε θα είναι γραμμικώς συσχετισμένες ανεξάρτητα της δομικής διάστασης του προβλήματος.

Έστω ότι έχουμε ένα πρόβλημα παλινδρόμησης με p ανεξάρτητες μεταβλητές που αποτελούν το διάνυσμα $\mathbf{x} = (x_1, \dots, x_p)^T$. Αν το πρόβλημα έχει 1D δομή, τότε για κάποιο διάνυσμα $\beta = (\beta_1, \dots, \beta_p)^T$, η κατανομή της $y | \mathbf{x}$ εξαρτάται από το \mathbf{x} μόνο μέσω του γραμμικού μετασχηματισμού $\beta^T \mathbf{x}$. Έστω ότι έχουμε το μοντέλο (2.1), όπου τόσο η συνάρτηση παλινδρόμησης όσο και η συνάρτηση διακύμανσης εξαρτώνται από το $\beta^T \mathbf{x}$. Ένα επαρκές διάγραμμα περίληψης θα είναι το $\{\beta^T \mathbf{x}, y\}$ αλλά αυτό απαιτεί τη γνώση του διανύσματος β , το οποίο μπορούμε να εκτιμήσουμε. Από το διάγραμμα αυτό μπορούμε να οπτικοποιήσουμε ταυτόχρονα και τη συνάρτηση f και τη συνάρτηση σ . Αν έχουμε γραμμικώς συσχετισμένες ανεξάρτητες μεταβλητές μπορούμε να πάρουμε χρήσιμη πληροφορία για το διάνυσμα β . Έστω ότι

$$\hat{y} = b_0 + b^T x,$$

όπου \hat{y} είναι οι προσαρμοσμένες τιμές μέσω της μεθόδου των ελαχίστων τετραγώνων. Για να κάνουμε αυτή την παλινδρόμηση, δεν υποθέτουμε την ισχύ του πολυμεταβλητού γραμμικού μοντέλου.

Υποθέτοντας γραμμικώς συσχετισμένες μεταβλητές, το $\hat{\mathbf{b}}$ είναι μια εκτίμηση του β για κάποια σταθερά c . Καθώς δεν μας ενδιαφέρει το μέγεθος των στοιχείων του β , το διδιάστατο γράφημα $\{\hat{\mathbf{b}}^T \mathbf{x}, y\}$ είναι μια εκτίμηση του επαρκούς διαγράμματος περίληψης $\{\beta^T \mathbf{x}, y\}$. Ισοδύναμα μπορούμε να πάρουμε το διάγραμμα $\{\hat{y}, y\}$ σαν διάγραμμα περίληψης. Αν το πραγματικό μοντέλο είναι το δομής 1D μοντέλο (2.1), τότε το διάγραμμα περίληψης μας επιτρέπει να οπτικοποιήσουμε τις συναρτήσεις f και σ . Το αποτέλεσμα αυτό καλείται *αποτέλεσμα εκτίμησης 1D* (1D estimation result) και είναι χρήσιμο ακόμα και όταν η δομική διάσταση της παλινδρόμησης είναι μεγαλύτερη του ένα. Ο γραμμικός συνδυασμός $\hat{\mathbf{b}}^T \mathbf{x}$ θα είναι πλέον ένας από τους γραμμικούς συνδυασμούς που απαιτούνται, και έτσι το διάγραμμα $\{\hat{y}, y\}$ θα είναι ακόμη σχετικό αν και θα χάνει σε πληροφορία.

Όταν έχουμε πολλές ανεξάρτητες μεταβλητές, χρειαζόμαστε δυο υποθέσεις. Αυτή της γραμμικής συσχέτισης των ανεξάρτητων μεταβλητών και της 1D δομής. Αυτά δεν χρειάζονται στην περίπτωση των δυο ανεξάρτητων μεταβλητών. Μπορούμε να πούμε ότι οι δυο αυτές υποθέσεις είναι το κόστος που πληρώνουμε γιατί δεν μπορούμε να δούμε πολλές διαστάσεις μαζί.

Συνοψίζοντας, να πούμε ότι η μέθοδος ελαχίστων τετραγώνων όταν έχουμε p ανεξάρτητες μεταβλητές, βρίσκει τη διδιάστατη απεικόνιση $\{\hat{y}, y\}$ ή $\{\hat{\mathbf{b}}^T \mathbf{x}, y\}$ με την ισχυρότερη γραμμική τάση. Ο τρόπος με τον οποίο ερμηνεύουμε την απεικόνιση εξαρτάται από τη δομή των δεδομένων. Έτσι, αν ένα γραμμικό μοντέλο της μορφής $y | \mathbf{x} = \beta_0 + \beta^T \mathbf{x} + \varepsilon$, είναι κατάλληλο, τότε δεν μας ενδιαφέρει η κατανομή των ανεξάρτητων τυχαίων μεταβλητών και το $\{\hat{y}, y\}$ είναι ένα καλό διάγραμμα περίληψης. Εάν η συνάρτηση παλινδρόμησης είναι μη γραμμική ως προς το διάνυσμα \mathbf{x} αλλά έχει 1D δομή και γραμμικές ανεξάρτητες μεταβλητές, τότε το $\{\hat{y}, y\}$ είναι πάλι καλό διάγραμμα. Εάν η συνάρτηση παλινδρόμησης είναι μη γραμμική ως προς το \mathbf{x} με μη γραμμικές ανεξάρτητες μεταβλητές, τότε το $\{\hat{y}, y\}$ δεν θεωρείται

καλό ακόμη και αν το πραγματικό μοντέλο έχει 1D δομή. Εάν, τέλος, το μοντέλο έχει μεγαλύτερη της 1D δομής, τότε το διάγραμμα $\{\hat{y}, y\}$ υποχρεωτικά χάνει πληροφορία που μας ενδιαφέρει (Cook and Weisberg, 1994).

2.2 Εύρεση της διάστασης της παλινδρόμησης

Στην προηγούμενη ενότητα είδαμε ότι το διδιάστατο διάγραμμα $\{\hat{y}, y\}$ συνοψίζει ικανοποιητικά το πρόβλημα της παλινδρόμησης, στην περίπτωση που έχουμε γραμμικώς συσχετισμένες ανεξάρτητες μεταβλητές και 1D δομή. Η γραμμικότητα των ανεξάρτητων μεταβλητών μπορεί να ελεγχθεί μέσω των πινάκων των διαγραμμάτων διασποράς ενώ στην περίπτωση που η γραμμικότητα δεν ισχύει, μπορούμε να την επιτύχουμε μέσω μετασχηματισμών. Στην ενότητα αυτή θα περιγράψουμε δυο μεθόδους για να ελέγχουμε την υπόθεση της 1D δομής όταν έχουμε περισσότερες των δυο ανεξάρτητες μεταβλητές. Η πρώτη είναι η γραφική μέθοδος και η δεύτερη που θα μας απασχολήσει περισσότερο είναι η μέθοδος της τμηματικής αντίστροφης παλινδρόμησης (sliced inverse regression - SIR).

2.2.1 Γραφική μέθοδος εύρεσης των διαστάσεων

Υποθέτουμε αρχικά ότι ισχύει η υπόθεση της 1D δομής. Για να ελέγξουμε την ανεξαρτησία των ερμηνευτικών μεταβλητών και να επιβεβαιώσουμε ότι ισχύει η υπόθεση της 1D δομής, θα πρέπει να κατασκευάσουμε ένα $(p+1)$ -διάστατο γράφημα των p ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής. Αυτό βέβαια δεν είναι εύκολο όταν ο αριθμός των ανεξάρτητων μεταβλητών είναι μεγαλύτερος του δυο. Υποθέτουμε ότι μπορούμε να αντιστρέψουμε το πρόβλημα και αντί να μελετήσουμε την κατανομή της $y|x$ να μελετήσουμε την αντίστροφη παλινδρόμηση (inverse regression), να δούμε δηλαδή πώς συμπεριφέρεται η κατανομή του διανύσματος x δοθεισών των τιμών της ανεξάρτητης μεταβλητής y , δηλαδή η κατανομή του $x|y$. Έτσι έχουμε να

αντιμετωπίσουμε p προβλήματα απλής παλινδρόμησης που μπορούν να μελετηθούν με ένα απλό διδιάστατο διάγραμμα διασποράς το καθένα.

Υποθέτουμε στη συνέχεια ότι έχουμε γραμμικές ανεξάρτητες μεταβλητές και 1D δομή. Οι συναρτήσεις παλινδρόμησης και διακύμανσης για καθένα από τα απλά προβλήματα παλινδρόμησης είναι αντίστοιχα

$$E(x_j | y) = E(x_j) + a_j m(y) \quad (2.5)$$

και

$$\text{var}(x_j | y) \approx \text{var}(x_j) + \alpha_j^2 v(y) \quad (2.6)$$

για $j = 1, \dots, p$, όπου τα a_j είναι ίδια για όλες τις εξισώσεις και μπορούν να πάρουν οποιαδήποτε θετική, αρνητική ή μηδενική τιμή.

Τα διαγράμματα $\{y, x_j\}$, $j = 1, \dots, p$, ονομάζονται αντίστροφα διαγράμματα μερικής απόκρισης (inverse partial response plots). Σημαντικό σημείο στην εξίσωση (2.5) είναι ότι η συνάρτηση $m(y)$ δεν εξαρτάται από το j . Αν έχουμε 1D δομή τα p αντίστροφα διαγράμματα μερικής απόκρισης θα πρέπει να έχουν την ίδια μορφή. Δηλαδή αν ένα από αυτά τα διαγράμματα είναι γραμμικό τότε όλα θα πρέπει να είναι γραμμικά. Αν ένα από αυτά έχει μορφή J τότε όλα θα πρέπει να έχουν μορφή J διαφέροντας μόνο σε μια σταθερά που μπορεί να αλλάζει τον προσανατολισμό. Αν κάτι τέτοιο δεν ισχύει, τότε η υπόθεση της 1D δομής θα πρέπει να εγκαταλειφθεί.

Στη σχέση (2.6) παρατηρούμε ότι η συνάρτηση $v(y)$ εξαρτάται από το y αλλά όχι από το j . Αυτό σημαίνει ότι παρατηρώντας τα διαγράμματα, θα πρέπει να παρατηρούμε ότι η διακύμανση αλλάζει με τον ίδιο τρόπο. Αν αυτό δεν συμβαίνει τότε δεν ισχύει η υπόθεση της 1D δομής. Η μόνη εξαίρεση είναι όταν έχουμε ανεξάρτητες μεταβλητές για τις οποίες τα a_j στις σχέσεις (2.5) και (2.6) είναι ίσα με μηδέν. Για να είναι συνεπή με την 1D δομή, τα διαγράμματα που δείχνουν μη εξάρτηση της συνάρτησης αντίστροφης παλινδρόμησης από το y , δηλαδή ισχύει $a_j = 0$, θα πρέπει να δείχνουν ανεξαρτησία και της συνάρτησης αντίστροφης διακύμανσης ακόμη και αν η διακύμανση δεν είναι σταθερή σε άλλα διαγράμματα. Οι εξισώσεις (2.5) και (2.6) ονομάζονται συνθήκες ελέγχου για την 1D δομή, καθώς οι εξισώσεις αυτές θα πρέπει να ισχύουν για κάθε εξαρτημένη μεταβλητή έτσι ώστε να ισχύει η 1D δομή. Στην περίπτωση που οι ανεξάρτητες μεταβλητές

ακολουθούν την κανονική κατανομή, τότε στη σχέση (2.6) ισχύει η ισότητα (Cook and Weisberg, 1999).

Σε καθένα από τα p αντίστροφα διαγράμματα μερικής απόκρισης, μπορούμε να προσαρμόσουμε έναν εξομαλυντή (smoother). Η ομαλή συνάρτηση που προσαρμόζεται στο j γράφημα, αποτελεί εκτίμηση της συνάρτησης αντίστροφης παλινδρόμησης $E(x_j | y)$. Σύμφωνα με την εξίσωση (2.5), αν ισχύει η 1D δομή, η συνάρτηση αντίστροφης παλινδρόμησης θα πρέπει να προσεγγίζει την ποσότητα $E(x_j) + a_j m(y)$. Σε κάθε διάγραμμα μπορεί να εφαρμοστεί διαφορετικός εξομαλυντής (Cook and Weisberg, 1994).

Ο Li (1991), θεωρώντας την καμπύλη που δημιουργείται από την $E(\mathbf{x} | y)$ καθώς η y μεταβάλλεται, την οποία ονομάζει ως καμπύλη αντίστροφης παλινδρόμησης (inverse regression curve), αναφέρει ότι το κέντρο της βρίσκεται στο σημείο $E(E(\mathbf{x} | y)) = E(\mathbf{x})$. Όταν για κάποιο διάνυσμα \mathbf{b} , η υπό συνθήκη μέση τιμή $E(\mathbf{b}^T \mathbf{x} | \boldsymbol{\beta}_1^T \mathbf{x}, \dots, \boldsymbol{\beta}_K^T \mathbf{x})$ είναι γραμμική ως προς τους γραμμικούς συνδυασμούς $\boldsymbol{\beta}_1^T \mathbf{x}, \dots, \boldsymbol{\beta}_K^T \mathbf{x}$, τότε η κεντραρισμένη καμπύλη αντίστροφης παλινδρόμησης, η οποία είναι μια p -διάστατη καμπύλη, θα βρίσκεται σε έναν K -διάστατο υποχώρο. Η παραπάνω συνθήκη της γραμμικότητας της $E(\mathbf{b}^T \mathbf{x} | \boldsymbol{\beta}_1^T \mathbf{x}, \dots, \boldsymbol{\beta}_K^T \mathbf{x})$, ικανοποιείται όταν η κατανομή του διανύσματος \mathbf{x} είναι ελλειπτικά συμμετρική όπως είναι η πολυμεταβλητή κανονική κατανομή. Η υπόθεση αυτή φαίνεται να επιβάλλει αυστηρές απαιτήσεις για την κατανομή του \mathbf{x} . Αυτό σημαίνει ότι κατά τη συλλογή των δεδομένων, θα πρέπει το πείραμα να έχει σχεδιαστεί έτσι ώστε η κατανομή του \mathbf{x} να μην παραβιάζει κατά πολύ την ελλειπτική συμμετρία.

Μια ενδιαφέρουσα επέκταση της υπόθεσης είναι η ποσοτικοποίηση του κατά πόσο διαφέρει η καμπύλη αντίστροφης παλινδρόμησης $E(\mathbf{z} | y)$, όπου $\mathbf{z} = \Sigma_{\mathbf{xx}}^{-1/2} (\mathbf{x} - E(\mathbf{x}))$ το διάνυσμα με τα τυποποιημένα στοιχεία του διανύσματος \mathbf{x} και $\Sigma_{\mathbf{xx}}$ ο πίνακας διακυμάνσεων του \mathbf{x} , από τον τυποποιημένο χώρο αποτελεσματικής μείωσης των διαστάσεων (effective dimension-reduction space). Ο Li (1991), αναφέρει κάθε γραμμικό συνδυασμό των διανυσμάτων $\boldsymbol{\beta}$ ως κατεύθυνση αποτελεσματικής μείωσης των

διαστάσεων (effective dimension-reduction direction) και το γραμμικό χώρο που παράγεται από τα β ως χώρο αποτελεσματικής μείωσης των διαστάσεων.

Στη συνέχεια, θα προσπαθήσουμε να βρούμε τη δομική διάσταση ενός προβλήματος παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής *BigMac* και των μετασχηματισμένων μεταβλητών $\log(BusFare)$, $\log(TeachTax)$, $\log(TeachSal)$ και $\log(Bread)$, χωρίς να υποθέτουμε την ισχύ κάποιου συγκεκριμένου μοντέλου με τη γραφική μέθοδο. Τα δεδομένα αυτά περιέχονται στο αρχείο «*big-mac.lsp*», το οποίο επίσης μπορούμε να κατεβάσουμε από τη διεύθυνση <http://www.stat.umn.edu/arc/>.

Σύντομη περιγραφή των μεταβλητών υπάρχει στον Πίνακα 2. Το πρόβλημα αυτό, επειδή έχει τέσσερις ανεξάρτητες μεταβλητές μπορεί να έχει μέχρι 4D δομή.

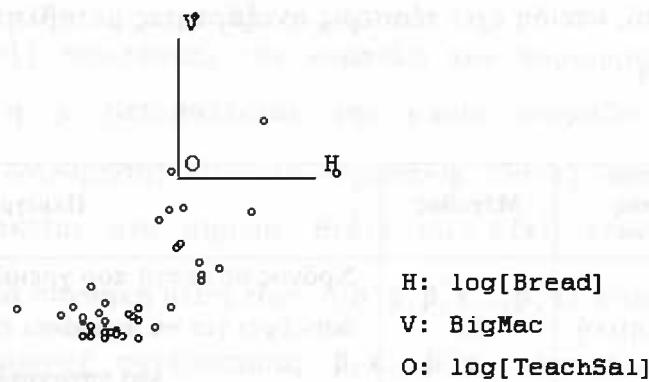
Όνομα μεταβλητής	Τύπος	Μέγεθος	Περιγραφή
BigMac	Αριθμητική	45	Χρόνος σε λεπτά που χρειάζεται ο μέσος εργάτης να δουλέψει για να αγοράσει ένα BigMac χάμπουργκερ και τηγανητές πατάτες
Bread	Αριθμητική	45	Χρόνος σε λεπτά που χρειάζεται ο μέσος εργάτης να δουλέψει για να αγοράσει ένα κιλό ψωμί
BusFare	Αριθμητική	45	Ελάχιστη τιμή εισιτηρίου για μια διαδρομή 10χλμ με δημόσια συγκοινωνία σε δολάρια
TeachSal	Αριθμητική	45	Μέσος ετήσιος μισθός ενός δασκάλου, σε χιλιάδες δολαρίων
TeachTax	Αριθμητική	45	Ποσοστό φόρου που πληρώνει ο μέσος δάσκαλος

Πίνακας 2.1: Περιγραφή μεταβλητών προβλήματος *BigMac*

Περιστρέφοντας και παρατηρώντας το τρισδιάστατο διάγραμμα του Σχήματος 2.3, δηλαδή το διάγραμμα $\{\log(Bread), \log(TrainFare), \log(TeachSal)\}$ συμπεραίνουμε ότι αυτό παρουσιάζει κάποιο συστηματικό πρότυπο. Συγκεκριμένα φαίνεται ότι οι παρατηρήσεις είναι πιο συγκεντρωμένες κάτω δεξιά από ότι είναι πάνω αριστερά. Αυτό, αν και μπορεί να οφείλεται στο ότι ο κύριος όγκος των παρατηρήσεων βρίσκεται κάτω δεξιά, σημαίνει ότι έχουμε αυξανόμενη διακύμανση. Αυτό οδηγεί στο συμπέρασμα ότι πρέπει να

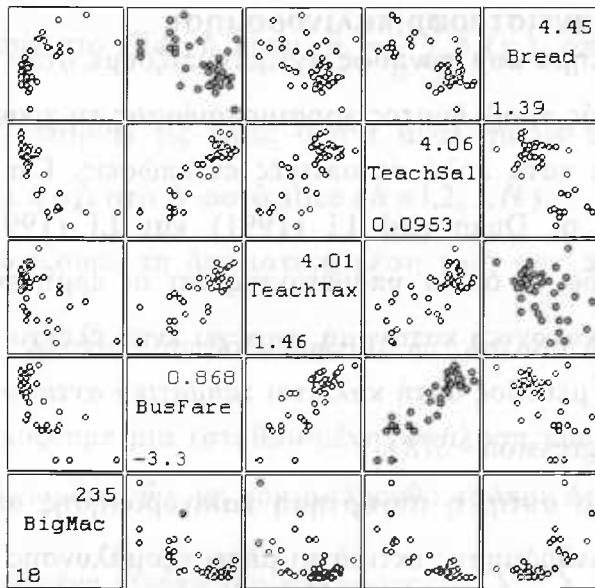
απορρίψουμε την υπόθεση της μηδενικής δομικής διάστασης. Συνεπώς το διάγραμμα θα έχει είτε 1D είτε 2D δομή. Για να βρούμε τι είδους δομή έχει θα χρησιμοποιήσουμε τη μέθοδο με την ασυσχέτιστη απεικόνιση και το slicer, που έχουμε περιγράψει στο Κεφάλαιο 1. Επειδή τα σημεία εμφανίζονται να βρίσκονται σε μια οριζόντια ζώνη, θεωρούμε ότι η μεταβλητή *BigMac* εξαρτάται από ένα μόνο γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών, οπότε το διάγραμμα έχει 1D δομή. Ο γραμμικός αυτός συνδυασμός είναι ο

$$h^* \approx -0.27 \log(Bread) + 0.38 \log(TrainSal).$$

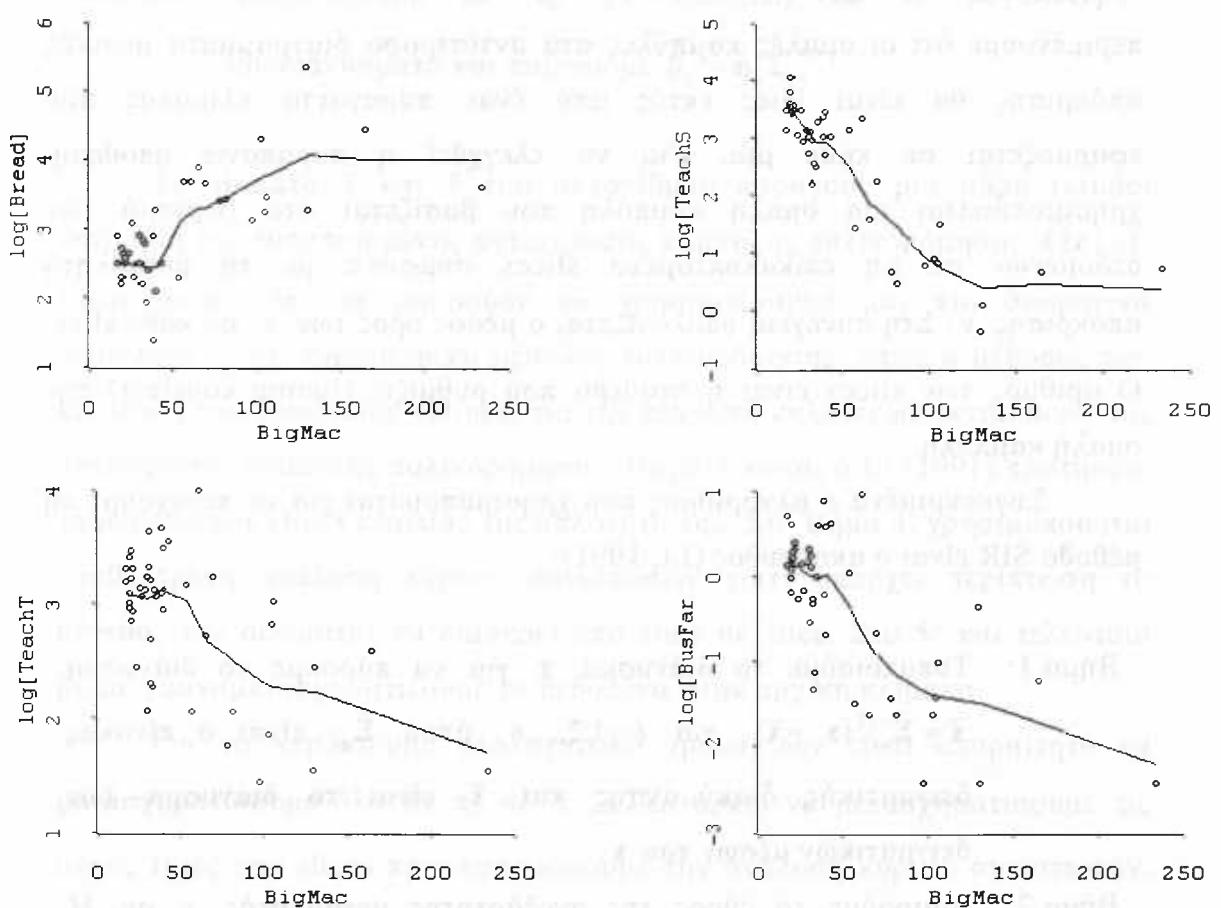


Σχήμα 2.3: Τρισδιάστατο διάγραμμα μεταξύ των μεταβλητών $\log(Bread)$, $BigMac$ και $\log(TrainSal)$

Στη συνέχεια θα χρησιμοποιήσουμε και τις τέσσερις μετασχηματισμένες μεταβλητές $\log(BusFare)$, $\log(TrainTax)$, $\log(TrainSal)$ και $\log(Bread)$. Από τον πίνακα διαγραμμάτων διασποράς του Σχήματος 2.4, συμπεραίνουμε ότι εκτός από μερικά σημεία ισχύει η υπόθεση των γραμμικών ανεξάρτητων μεταβλητών. Από την πρώτη στήλη του πίνακα, που παρουσιάζει τα διαγράμματα αντίστροφης παλινδρόμησης, συμπεραίνουμε ότι επειδή και τα τέσσερα διαγράμματα δείχνουν περίπου το ίδιο πρότυπο, το πρόβλημα πρέπει να έχει 1D δομή. Τα τέσσερα διαγράμματα μερικής απόκρισης, δίνονται στο Σχήμα 2.5. Σε κάθε διάγραμμα έχουμε προσαρμόσει μια ομαλή καμπύλη *lowess* με παράμετρο 0.7.



Σχήμα 2.4: Πίνακας διαγραμμάτων διασποράς με μετασχηματισμένες σε λογαριθμική κλίμακα ανεξάρτητες μεταβλητές



Σχήμα 2.5: Διαγράμματα αντίστροφης παλινδρόμησης με προσαρμοσμένη *lowess*

2.3 Τμηματική αντίστροφη παλινδρόμηση

Ένα πρόβλημα που συνήθως αντιμετωπίζουμε όταν αποφασίζουμε για τις διαστάσεις ενός προβλήματος χρησιμοποιώντας τη γραφική μέθοδο, είναι ότι στηριζόμαστε κατά πολύ σε οπτικές εντυπώσεις. Για να ξεπεραστεί το πρόβλημα αυτό, οι Duan and Li (1991) και Li (1991) πρότειναν μια αριθμητική μέθοδο, η οποία υποθέτοντας ότι οι ερμηνευτικές μεταβλητές ακολουθούν την κανονική κατανομή, παρέχει έναν έλεγχο για την ισχύ ή όχι της 1D δομής. Η μέθοδος αυτή καλείται *τμηματική αντίστροφη παλινδρόμηση* (sliced inverse regression - *SIR*).

Στη μέθοδο αυτή, η συνάρτηση παλινδρόμησης σε κάθε διάγραμμα αντίστροφης παλινδρόμησης εκτιμάται μέσω εξομάλυνσης (smoothing). Ο Li (1991) έδειξε ότι οι εκτιμημένες ομαλές καμπύλες μπορούν να συγκριθούν για να δώσουν ένα τεστ για τη διάσταση του προβλήματος. Υποθέτουμε ότι υπάρχει μια συνάρτηση $m(y)$ και ότι ισχύει η εξίσωση (2.5). Τότε περιμένουμε ότι οι ομαλές καμπύλες στα αντίστροφα διαγράμματα μερικής απόκρισης θα είναι ίδιες εκτός από έναν παράγοντα κλίμακας που εφαρμόζεται σε κάθε μια. Για να ελεγχθεί η παραπάνω υπόθεση, χρησιμοποιείται μια ομαλή καμπύλη που βασίζεται στο χωρισμό των δεδομένων σε μη επικαλυπτόμενα slices σύμφωνα με τη μεταβλητή απόκρισης y . Στη συνέχεια υπολογίζεται ο μέσος όρος των x_i σε κάθε slice. Ο αριθμός των slices είναι η σταθερά που ρυθμίζει (tuning constant) την ομαλή καμπύλη.

Συγκεκριμένα ο αλγόριθμος που χρησιμοποιείται για να πετύχουμε τη μέθοδο SIR είναι ο ακόλουθος (Li, 1991):

Βήμα 1: Τυποποιούμε το διάνυσμα \mathbf{x} για να πάρουμε το διάνυσμα

$$\tilde{\mathbf{x}} = \hat{\Sigma}_{\mathbf{xx}}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), \text{ για } i = 1, 2, \dots, n, \text{ όπου } \hat{\Sigma}_{\mathbf{xx}} \text{ είναι ο πίνακας δειγματικής διακύμανσης και } \bar{\mathbf{x}} \text{ είναι το διάνυσμα των δειγματικών μέσων του } \mathbf{x}.$$

Βήμα 2: Διαιρούμε το εύρος της ανεξάρτητης μεταβλητής y σε H slices, τις I_1, \dots, I_H . Έστω ότι το ποσοστό των τιμών y_i που

ανήκουν στο slice h είναι $\hat{p}_h = \frac{1}{n} \sum_{i=1}^n \delta_h(y_i)$, όπου η συνάρτηση

$\delta_h(y_i)$ παίρνει τις τιμές 0 ή 1 ανάλογα με το αν η τιμή y_i ανήκει ή όχι στο h -οστό slice ($h = 1, 2, \dots, H$).

Βήμα 3: Υπολογίζουμε τη δειγματική μέση τιμή των \mathbf{x} , μέσα σε κάθε

slice, την οποία συμβολίζουμε με \hat{m}_h , δηλαδή $\hat{m}_h = \frac{1}{n\hat{p}_h} \sum_{i \in I_h} \tilde{x}_i$.

Βήμα 4: Εφαρμόζουμε μια (σταθμισμένη) ανάλυση κύριων συνιστωσών

στα δεδομένα \hat{m}_h με τον ακόλουθο τρόπο: Δημιουργούμε τον σταθμισμένο πίνακα συνδιακυμάνσεων $\hat{\mathbf{V}} = \sum_{h=1}^H \hat{p}_h \hat{m}_h \hat{m}_h^T$ και στη

συνέχεια βρίσκουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του $\hat{\mathbf{V}}$.

Βήμα 5: Συμβολίζουμε με $\hat{\boldsymbol{\eta}}_k$ ($k = 1, 2, \dots, K$) τα K μεγαλύτερα ιδιοδιανύσματα και παίρνουμε $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\eta}}_k \hat{\Sigma}_{\mathbf{xx}}^{-1/2}$

Τα Βήματα 2 και 3 του αλγορίθμου παράγουν μια αδρή (crude) εκτίμηση της τυποποιημένης αντίστροφης καμπύλης παλινδρόμησης $E(\mathbf{z} | y)$. Είναι σαφές ότι θα μπορούσε να χρησιμοποιηθεί μια πιο θεωρητικά θεμελιωμένη μη παραμετρική μέθοδος παλινδρόμησης, όπως η μέθοδος των kernels ή των smoothing splines, για την εξαγωγή καλύτερων εκτιμήσεων της αντίστροφης καμπύλης παλινδρόμησης. Παρόλα αυτά, ο Li (1991) προτίμησε τη μέθοδο των slices εξαιτίας της απλότητά της. Στο Βήμα 4, χρησιμοποιείται σταθμισμένη ανάλυση κύριων συνιστωσών γιατί υπάρχει περίπτωση το μέγεθος του δείγματος να διαφέρει από slice σε slice. Στο 5^o και τελευταίο βήμα, επαναμετασχηματίζουμε τα δεδομένα στην αρχική κλίμακα.

Για να κερδίσουμε υπολογιστικό χρόνο, δεν είναι απαραίτητο να μετασχηματίσουμε όλα τα \mathbf{x} , σε $\tilde{\mathbf{x}}$, αλλά αρκεί να μετασχηματίσουμε τις μέσες τιμές των slices πριν εφαρμόσουμε την ανάλυση κύριων συνιστωσών.

Συμβολίζουμε με $\hat{\Sigma}_1$ τον πίνακα $\sum_{h=1}^H \hat{p}_h (\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})^T$, όπου με $\bar{\mathbf{x}}_h$

συμβολίζουμε το δειγματικό μέσο των \mathbf{x}_i στο h -οστό slice. Τότε τα $\hat{\boldsymbol{\beta}}_k$ είναι

τα ιδιοδιανύσματα για την eigenvalue decomposition του πίνακα $\hat{\Sigma}_1$. Επίσης μπορεί να οριστεί η μέθοδος, έτσι ώστε κάθε slice να έχει το ίδιο εύρος, χωρίς αυτό να έχει καμιά ιδιαίτερη σημασία. Η επιλογή του αριθμού των slices μπορεί να επηρεάσει την ασυμπτωτική διακύμανση των εκτιμήσεων. Παρόλα αυτά, η προσομοίωση που έκανε ο Li (1991), έδειξε ότι η διαφορά στην ασυμπτωτική διακύμανση δεν είναι σημαντική για μεγέθη δείγματος που χρησιμοποιούνται στην πράξη.

Ακολουθώντας την παραπάνω διαδικασία, κατασκευάζεται ένα τρισδιάστατο γράφημα της μορφής $\{h_1, y, h_2\}$, όπου τα $h_1 = \mathbf{b}^T \mathbf{x}$ και $h_2 = \mathbf{a}^T \mathbf{x}$ είναι γραμμικοί συνδυασμοί των ανεξάρτητων μεταβλητών που έχουν εκτιμηθεί με την παραπάνω μέθοδο. Αν η 1D δομή είναι πραγματικά κατάλληλη, τότε η διδιάστατη απεικόνιση $\{h_1, y\}$ πριν από την περιστροφή είναι το διάγραμμα περίληψης που εκτιμάται από τη μέθοδο *SIR*. Αν χρειάζεται δομή 2D τότε απαιτείται και δεύτερος γραμμικός συνδυασμός, ο οποίος είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών h_2 στον εκτός σελίδας άξονα. Αν ισχύει η 2D δομή, τότε το πλήρες τρισδιάστατο γράφημα $\{h_1, y, h_2\}$ αποτελεί το διάγραμμα περίληψης που προκύπτει από τη μέθοδο τμηματικής αντίστροφης παλινδρόμησης. Εκ κατασκευής, οι γραμμικοί συνδυασμοί h_1 και h_2 είναι μεταξύ τους ασυσχέτιστοι.

Αν εφαρμόσουμε τη μέθοδο τμηματικής αντίστροφης παλινδρόμησης με το πρόγραμμα *Arc*, παίρνουμε ένα output σαν αυτό που εμφανίζεται στον Πίνακα 2.2. Για την παραγωγή του output αυτού, χρησιμοποιήθηκε και πάλι το set δεδομένων «*big-mac.lsp*», χωρίς να έχουμε κάνει καμιά μετατροπή στα δεδομένα.

Παρατηρώντας τα αποτελέσματα, βλέπουμε ότι στην αρχή, δίνονται τα διανύσματα των συντελεστών **b** και **a** (υπό τη στήλη με όνομα **Raw**) που αντιστοιχούν στους γραμμικούς συνδυασμούς h_1 και h_2 . Σημειώνουμε ότι και τα δύο διανύσματα είναι κανονικοποιημένα έτσι ώστε να έχουν μήκος 1. Οι στήλες με όνομα **Std.** δίνουν τους συντελεστές που θα είχαμε πάρει αν κάθε μια από τις εξαρτημένες μεταβλητές είχε επανακλιμακωθεί έτσι ώστε να έχει τυπική απόκλιση ίση με τη μονάδα. Και οι τυποποιημένοι συντελεστές είναι κανονικοποιημένοι έτσι ώστε τα διανύσματά τους να έχουν μήκος 1.

Κάτω από τα διανύσματα των συντελεστών, υπάρχουν δυο γραμμές που αναφέρονται ως **Eigenvalues και R^2 (OLS|SIR lin comb)**, αντίστοιχα. Σημειώνουμε ότι πάντα ο πρώτος γραμμικός συνδυασμός θα έχει τη μεγαλύτερη ιδιοτιμή (eigenvalue). Ο έλεγχος που δίνει η μέθοδος για τη διαστασιμότητα του προβλήματος, είναι συνάρτηση των ιδιοτιμών. Το μέτρο R^2 αποτελεί μια περίληψη της συμφωνίας μεταξύ των γραμμικών συνδυασμών h_1 και h_2 , που έχουν επιλεγεί από τη μέθοδο και των προσαρμοσμένων μέσω της μεθόδου των ελαχίστων τετραγώνων τιμών \hat{y} . Είναι σαφές ότι όσο πιο μεγάλη η τιμή του R^2 τόσο καλύτερα. Σύμφωνα με τον Li (1991), το R^2 είναι συνάρτηση του διανύσματος \mathbf{b} και ισούται με το τετράγωνο του συντελεστή πολλαπλής παλινδρόμησης μεταξύ του γραμμικού συνδυασμού $\mathbf{b}^T \mathbf{x}$ και των ιδεατά μειωμένων μεταβλητών $\mathbf{b}_1^T \mathbf{x}, \dots, \mathbf{b}_K^T \mathbf{x}$, ισχύει δηλαδή η σχέση

$$R^2(\mathbf{b}) = \max_{\mathbf{b} \in \mathcal{B}} \frac{(\mathbf{b} \Sigma_{xx} \boldsymbol{\beta}^T)^2}{\mathbf{b} \Sigma_{xx} \mathbf{b}^T \cdot \boldsymbol{\beta} \Sigma_{xx} \boldsymbol{\beta}^T}.$$

Σημειώνουμε ότι το διάνυσμα \mathbf{b} είναι κανονικοποιήμενο έτσι ώστε να έχει μέγεθος 1, ισχύει δηλαδή $\|\mathbf{b}\| = \sqrt{\mathbf{b}' \mathbf{b}} = 1$.

Η μέθοδος τμηματικής αντίστροφης παλινδρόμησης παρέχει μια σειρά από στατιστικά τεστ που βοηθούν στην απόφαση για τις διαστάσεις του προβλήματος της παλινδρόμησης. Συγκεκριμένα, παρέχονται τρεις έλεγχοι και κάθε στατιστικό συγκρίνεται με τα ποσοστιαία σημεία μιας χ^2 κατανομής με κατάλληλους βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας της κατανομής εξαρτώνται από τον αριθμό των slices που χρησιμοποιούνται για να πάρουμε τις ομαλές καμπύλες στα διαγράμματα διασποράς.

Inverse Regression SIR, Name of Dataset = Mac

Response = BigMac

Predictors = (Bread TeachSal TeachTax BusFare WorkHrs)

Number of slices = 11

Slices sizes are: (6 5 4 4 4 4 4 4 3 3)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2		Lin Comb 3	
Predictors	Raw	Std.	Raw	Std.	Raw	Std.
Bread	-0.030	-0.263	-0.084	-0.148	-0.012	-0.141
TeachSal	0.238	0.901	-0.475	-0.366	0.075	0.394
TeachTax	-0.105	-0.286	-0.326	-0.181	-0.226	-0.852
BusFare	0.965	0.162	-0.807	-0.028	-0.971	-0.224
WorkHrs	-0.002	-0.102	-0.098	-0.901	-0.004	-0.219
Eigenvalues	0.812		0.381		0.324	
R^2(OLS SIR)	0.924		0.926		0.952	

Approximate Chi-squared test statistics based on partial sums of eigenvalues times 45

Number of Components	Test Statistic	df	p-value
1	85.161	50	0.001
2	48.604	36	0.078
3	31.469	24	0.141
4	16.9	14	0.262

Πίνακας 2.2: Αποτελέσματα μεθόδου SIR

Ο Li (1991), έδειξε ότι αν το διάνυσμα \mathbf{x} ακολουθεί την πολυμεταβλητή κανονική κατανομή, τότε η στατιστική συνάρτηση $n(p-K)\bar{\lambda}_{(p-K)}$, όπου $\bar{\lambda}_{(p-K)}$ είναι η μέση τιμή των $(p-K)$ μικρότερων ιδιοτιμών του πίνακα συνδιακυμάνσεων του πίνακα $\hat{\mathbf{V}}$, ακολουθεί ασυμπτωτικά την χ^2 κατανομή με $(p-K)(H-K-1)$ βαθμούς ελευθερίας.

Για άλλες, πλην της κανονικής, ελλειπτικά συμμετρικές κατανομές το αποτέλεσμα είναι πιο περίπλοκο.

Το πρώτο τεστ αφορά τον έλεγχο της υπόθεσης ότι η παλινδρόμηση έχει 0D δομή έναντι της εναλλακτικής ότι η δομή του προβλήματος είναι τουλάχιστον 1. Το δεύτερο τεστ ελέγχει την υπόθεση ότι η παλινδρόμηση έχει το πολύ 1D δομή έναντι της εναλλακτικής ότι η δομή του προβλήματος είναι τουλάχιστον 2. Τέλος, το τρίτο τεστ ελέγχει την υπόθεση ότι η παλινδρόμηση έχει το πολύ 2D δομή έναντι της εναλλακτικής ότι η δομική διάσταση του προβλήματος είναι τουλάχιστον 3. Σημειώνουμε ότι σε επίπεδο σημαντικότητας 5%, η μηδενική υπόθεση απορρίπτεται υπέρ της εναλλακτικής αν το *p-value* του ελέγχου είναι μικρότερο της τιμής 0.05. Οι έλεγχοι της μεθόδου *SIR* απαιτούν οι ανεξάρτητες μεταβλητές να ακολουθούν την κανονική κατανομή. Η υπόθεση αυτή είναι πιο περιοριστική από την υπόθεση των γραμμικώς συσχετισμένων μεταβλητών και τα αποτελέσματα των ελέγχων είναι ευαίσθητα σε περιπτώσεις μη κανονικότητας. Αυτό έχει ως αποτέλεσμα, οι έλεγχοι να χρησιμοποιούνται ως απλοί οδηγοί που επιβεβαιώνονται από τα διαγράμματα αντίστροφης μερικής απόκρισης. Αν για παράδειγμα οι έλεγχοι δείχνουν 2D δομή η οποία όμως δεν είναι εμφανής στο τρισδιάστατο διάγραμμα της μεθόδου *SIR*, τότε είναι προτιμότερο να πιστέψουμε αυτό που βλέπουμε στο διάγραμμα. Σημειώνουμε ότι η μέθοδος δεν θα δώσει τα αποτελέσματα των ελέγχων στην περίπτωση που ο αριθμός των slices είναι μικρότερος ή ίσος του αριθμού $p+1$, όπου p είναι ο αριθμός των ανεξάρτητων μεταβλητών, αν και οι συντελεστές της μεθόδου μπορεί να είναι χρήσιμοι.

Εφαρμόζοντας τη μέθοδο *SIR*, επιλέγοντας 15 slices, στο set δεδομένων «*big-mac.lsp*», παίρνουμε τα αποτελέσματα του Πίνακα 2.3. Παρατηρώντας τον πρώτο έλεγχο που δίνει η μέθοδος *SIR*, συμπεραίνουμε ότι το πρόβλημα έχει τουλάχιστον 1D δομή, αφού το *p-value* του ελέγχου είναι $0.010 < 0.05$. Από τον δεύτερο έλεγχο, που ελέγχει την υπόθεση ότι η παλινδρόμηση έχει το πολύ 1D δομή έναντι της εναλλακτικής υπόθεσης ότι έχει το λιγότερο 2D δομή, συμπεραίνουμε ότι επειδή το *p-value* ισούται με 0.473, που ξεπερνά την τιμή 0.05, η παλινδρόμηση έχει 1D δομή. Αυτό συμφωνεί με το συμπέρασμα που είχαμε βγάλει και με τη γραφική μέθοδο.

Άρα η εξαρτημένη μεταβλητή *BigMac* εξαρτάται από ένα μόνο γραμμικό συνδυασμό των μετασχηματισμένων μεταβλητών, δηλαδή τον

$$h = -0.237 \log(Bread) + 0.859 \log(TeachSal) - 0.447 \log(TeachTax) + 0.08 \log(BusFare).$$

Στη συνέχεια θα μετασχηματίσουμε σε λογαριθμική κλίμακα και την εξαρτημένη μεταβλητή *BigMac* και θα εφαρμόσουμε πάλι τη μέθοδο *SIR* για να δούμε αν αυτό θα αλλάξει τη δομική διάσταση του προβλήματος. Τα αποτελέσματα της μεθόδου εμφανίζονται στον Πίνακα 2.4. Παρατηρούμε ότι αντικαθιστώντας την εξαρτημένη μεταβλητή με τον λογάριθμό της, τα αποτελέσματα δεν αλλάζουν σχεδόν καθόλου. Και πάλι συμπεραίνουμε ότι η δομική διάσταση του προβλήματος της παλινδρόμησης έχει 1D δομή και ο γραμμικός συνδυασμός είναι ο ίδιος με την προηγούμενη περίπτωση. Αυτό έρχεται να επιβεβαιώσει τον ισχυρισμό των Cook and Weisber (1994) ότι μονότονοι μετασχηματισμοί της εξαρτημένης μεταβλητής δεν επηρεάζουν τη δομική διάσταση του προβλήματος.

Inverse Regression SIR, Name of Dataset = Mac

Response = BigMac

Predictors = (log[Bread] log[TeachSal] log[TeachTax] log[BusFare])

Number of slices = 13

Slices sizes are: (6 3 3 3 4 3 3 4 3 3 3 3 4)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2		Lin Comb 3	
Predictors	Raw	Std.	Raw	Std.	Raw	Std.
log[Bread]	-0.237	-0.168	0.601	0.615	-0.468	-0.438
log[TeachSal]	0.859	0.949	0.245	0.389	-0.478	-0.696
log[TeachTax]	-0.447	-0.260	0.707	0.591	0.743	0.568
log[BusFare]	0.080	0.069	-0.282	-0.349	-0.027	-0.030

Eigenvalues	0.906	0.547	0.127
R^2(OLS SIR)	0.983	0.983	0.991

Approximate Chi-squared test statistics based on partial sums of eigenvalues times 45

Number of Test

Components	Statistic	df	p-value
1	73.653	48	0.010
2	32.885	33	0.473
3	8.2476	20	0.990
4	2.5386	9	0.980

Πίνακας 2.3: Αποτελέσματα μεθόδου SIR με 15 slices

Inverse Regression SIR, Name of Dataset = Mac

Response = log[BigMac]

Predictors = (log[Bread] log[TeachSal] log[TeachTax] log[BusFare])

Number of slices = 13

Slices sizes are: (6 3 3 3 4 3 3 4 3 3 3 3 4)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2		Lin Comb 3	
Predictors	Raw	Std.	Raw	Std.	Raw	Std.
log[Bread]	-0.237	-0.168	0.601	0.615	-0.468	-0.438
log[TeachSal]	0.859	0.949	0.245	0.389	-0.478	-0.696
log[TeachTax]	-0.447	-0.260	0.707	0.591	0.743	0.568
log[BusFare]	0.080	0.069	-0.282	-0.349	-0.027	-0.030

Eigenvalues	0.906	0.547	0.127
R^2(OLS SIR)	0.996	0.998	0.998

Approximate Chi-squared test statistics based on partial sums of eigenvalues times 45

Number of Test

Components	Statistic	df	p-value
1	73.653	48	0.010
2	32.885	33	0.473
3	8.2476	20	0.990
4	2.5386	9	0.980

Πίνακας 2.4: Αποτελέσματα μεθόδου SIR με 15 slices και εξαρτημένη μεταβλητή την log(BigMac)

2.4 Μέθοδος SAVE

Οι Cook and Weisberg (1991), πρότειναν μια παρόμοια με τη μέθοδο SIR, την οποία ονόμασαν τμηματική μέση εκτίμηση της διακύμανσης (sliced average variance estimation - SAVE). Η μέθοδος αυτή χρησιμοποιεί τις υποθέσεις ελέγχου (2.5) και (2.6) που αναφέρονται στη συνάρτηση παλινδρόμησης και διακύμανσης αντίστοιχα για να βρει την αντίστροφη

δομή. Δηλαδή, η μέθοδος *SAVE* χρησιμοποιεί τόσο τη συνθήκη γραμμικότητας της συνάρτησης παλινδρόμησης όσο και τη συνθήκη της σταθερής συνάρτησης διακύμανσης. Σημειώνουμε ότι η μέθοδος *SIR* χρησιμοποιεί μόνο την υπόθεση που ελέγχει τη συνάρτηση παλινδρόμησης.

Οι Cook and Weisberg (1991), στο σχολιασμό της μεθόδου του Li (1991), αναφέρουν ότι το μοντέλο $y = f(\beta_1^T \mathbf{x}, \dots, \beta_k^T \mathbf{x}, \epsilon)$ που πρότεινε ο Li, δεν κάνει καμιά υπόθεση σχετικά με τη μορφή της εξάρτησης της εξαρτημένης μεταβλητής y από το διάνυσμα των ανεξάρτητων μεταβλητών \mathbf{x} . Η εξάρτηση μπορεί να είναι μέσω της συνάρτησης παλινδρόμησης, όπως αναφέρει ο Li, μπορεί όμως να είναι και μέσω της διακύμανσης ή κάποιας άλλης ανώτερης τάξης ροπής. Επίσης αναφέρουν ότι η υπόθεση της γραμμικότητας της υπό συνθήκη μέσης τιμής $E(\mathbf{b}^T \mathbf{x} | \beta_1^T \mathbf{x}, \dots, \beta_k^T \mathbf{x})$ είναι πολύ περιοριστική που σε ορισμένες περιπτώσεις είναι αμφίβολο αν μπορεί να εφαρμοσθεί. Για παράδειγμα, η συνθήκη αυτή αποκλείει κάποιους τυπικούς πειραματικούς σχεδιασμούς ή προβλήματα παλινδρόμησης με ψευδομεταβλητές (dummy variables).

Οι Cook and Weisberg (1991), εφάρμοσαν τη μέθοδο της τμηματικής αντίστροφης παλινδρόμησης σε ένα παράδειγμα και κατέληξαν στο συμπέρασμα ότι αυτή δεν καταλήγει πάντα σε σωστή απόφαση σχετικά με τις διαστάσεις σε ένα πρόβλημα παλινδρόμησης. Αυτό συμβαίνει όταν η διακύμανση των σφαλμάτων ϵ είναι μικρή. Στην περίπτωση που η διακύμανση των σφαλμάτων είναι μεγάλη η μέθοδος *SIR* βρίσκει πάντα τη σωστή λύση. Ένα δεύτερο μειονέκτημα της μεθόδου *SIR* είναι η ανικανότητά της στη διάγνωση της συμμετρικής εξάρτησης, όταν η μέση τιμή των τυποποιημένων τιμών του \mathbf{x} σε κάθε slice είναι κοντά στο μηδέν, όπως φαίνεται στο Σχήμα 2.6. Στην περίπτωση αυτή είναι πιθανόν όλες οι ιδιοτιμές του πίνακα συνδιακυμάνσεων που κατασκευάζεται από τα διανύσματα των μέσων τιμών των slices να έχουν το ίδιο μέγεθος. Αυτό σημαίνει ότι η εύρεση της κατεύθυνσης με τη μεγαλύτερη ιδιοτιμή που θα αντιστοιχηθεί στο διάνυσμα \mathbf{z}_1 θα γίνει κατά τύχη και είναι αδύνατη όταν ο αριθμός των ανεξάρτητων μεταβλητών ξεπερνά το 3 ή 4.

Από το Σχήμα 2.6 γίνεται εύκολα αντιληπτό, ότι μπορεί να ισχύει $E(\mathbf{z} | y) = 0$ αλλά η διακύμανση $\text{var}(\mathbf{z} | y)$ να μεταβάλλεται από slice σε slice.

Το γεγονός αυτό μπορεί να βοηθήσει ώστε να γίνει σωστή εκτίμηση της κατεύθυνσης αποτελεσματικής μείωσης των διαστάσεων, χρησιμοποιώντας δεύτερης ή μεγαλύτερης τάξης ροπές. Οι Cook and Weisberg (1991), θεωρώντας ότι η κατανομή του διανύσματος \mathbf{x} είναι ελλειπτικά συμμετρική και τυποποιώντας το διάνυσμα ώστε να πάρουν το \mathbf{z} , βρήκαν ότι ισχύει η σχέση

$$\text{var}(\mathbf{z} | y) = w_y Q_\eta + P_\eta \text{var}(\mathbf{z} | y) P_\eta^T$$

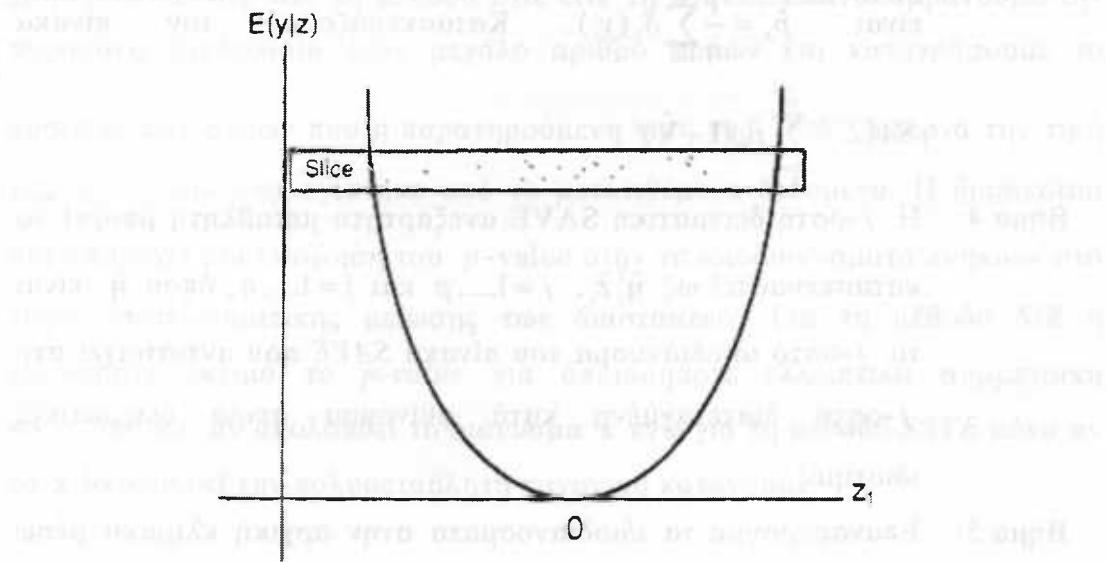
ή ισοδύναμα

$$w_y \mathbf{I} - \text{var}(\mathbf{z} | y) = P_\eta [w_y \mathbf{I} - \text{var}(\mathbf{z} | y)] P_\eta^T,$$

όπου Q_η και P_η είναι τελεστές προέκτασης (operation operators) για τους οποίους ισχύει $Q_\eta = \mathbf{I} - P_\eta$, και το w_y είναι μια συνάρτηση του y που εξαρτάται από την ελλειπτικά συμμετρική κατανομή του διανύσματος \mathbf{x} . Η συνάρτηση διακύμανσης $\text{var}(\mathbf{z} | y)$ έχει ιδιοτιμή w_y με πολλαπλασιαστικό παράγοντα (multiplicity) $p-K$ και τα αντίστοιχα ιδιοδιανύσματα διατρέχουν (span) τον υποχώρο αποτελεσματικής μείωσης των διαστάσεων του Q_η . Τα εναπομείναντα ιδιοδιανύσματα διατρέχουν τον υποχώρο αποτελεσματικής μείωσης των διαστάσεων. Καθώς $\text{var}(\mathbf{z}) = \mathbf{I}$ τότε $E(w_y) = 1$ και συνεπώς η συνάρτηση w_y θα διαφέρει από slice σε slice κατά μια μονάδα. Όταν η κατανομή του \mathbf{x} είναι η πολυμεταβλητή κανονική κατανομή, τότε ισχύει $w_y = 1$.

Το θέμα είναι πως μπορούν να συνδυαστούν οι πληροφορίες από το κάθε slice. Οι Cook and Weisberg (1991), ακολούθησαν τη λογική που είχε ακολουθήσει ο Li (1991). Δηλαδή είτε θα έβρισκαν τη μέση τιμή των υποχώρων που αντιστοιχούν στα επιλεγμένα ιδιοδιανύσματα σε κάθε slice είτε θα συνδύαζαν τις μεμονωμένες εκτιμήσεις των διακυμάνσεων $\text{var}(\mathbf{z} | y \in I_h)$, όπου I_h είναι η δείκτρια συνάρτηση για κάποιο slice, σε έναν πίνακα μέσω του οποίου θα εκτιμούσαν τον υποχώρο αποτελεσματικής μείωσης των διαστάσεων. Θεωρώντας slices με ίσα μεγέθη, εκτίμησαν τον υποχώρο αποτελεσματικής μείωσης των διαστάσεων χρησιμοποιώντας τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές του πίνακα

$$SAVE = \sum_h (\mathbf{I} - \text{var}(\mathbf{z} | y \in I_h))^2.$$



Σχήμα 2.6: Τμηματοποίηση όταν η y είναι τετραγωνική συνάρτηση ενός από τα \mathbf{z}

(Πηγή: Cook and Weisberg, 1991)

Η μέθοδος αυτή ονομάζεται εκτίμηση τμηματικής μέσης διακύμανσης (sliced average variance estimate - *SAVE*). Το κίνητρο για την χρησιμοποίηση του παραπάνω πίνακα προήλθε από τη σχέση

$$[\mathbf{I} - \text{var}(\mathbf{z} | y)]^2 = P_h [\mathbf{I} - \text{var}(\mathbf{z} | y)]^2 P_h^T$$

και από το γεγονός ότι οι ιδιοτιμές του πίνακα $[\mathbf{I} - \text{var}(\mathbf{z} | y)]^2$ δεν μπορεί να είναι αρνητικές.

Ο αλγόριθμος που χρησιμοποιείται για την επίτευξη της μεθόδου *SAVE* είναι ο ακόλουθος (Cook, 2000):

Βήμα 1: Τυποποιούμε το διάνυσμα \mathbf{x} για να πάρουμε το διάνυσμα $\mathbf{z} = \hat{\Sigma}_{\mathbf{xx}}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}})$, για $i = 1, 2, \dots, n$, όπου $\hat{\Sigma}_{\mathbf{xx}}$ είναι ο πίνακας δειγματικής διακύμανσης και $\bar{\mathbf{x}}$ είναι το διάνυσμα των δειγματικών μέσων του \mathbf{x} .

Βήμα 2: Διαιρούμε το εύρος της ανεξάρτητης y σε H slices, τις I_1, \dots, I_H και δημιουργούμε τον σταθμισμένο πίνακα

συνδιακυμάνσεων $\hat{\mathbf{V}}$ των \mathbf{z} .

Βήμα 3: Έστω ότι το ποσοστό των τιμών y_i που ανήκουν στο slice h

$$\text{είναι } \hat{p}_h = \frac{1}{n} \sum_{i=1}^n \delta_h(y_i). \quad \text{Κατασκευάζουμε τον πίνακα}$$

$$SAVE = \sum_{h=1}^H \hat{p}_h (\mathbf{I} - \hat{\mathbf{V}})^2.$$

Βήμα 4: Η j -οστή δειγματική $SAVE$ ανεξάρτητη μεταβλητή μπορεί να κατασκευαστεί ως $\hat{\boldsymbol{\eta}}_j^T \hat{\mathbf{z}}_i$, $j = 1, \dots, p$ και $i = 1, \dots, n$, όπου $\hat{\boldsymbol{\eta}}_j$ είναι το j -οστό ιδιοδιανύσμα του πίνακα $SAVE$ που αντιστοιχεί στη j -οστή διατεταγμένη κατά φθίνουσα σειρά δειγματικές ιδιοτιμές.

Βήμα 5: Επαναφέρουμε τα ιδιοδιανύσματα στην αρχική κλίμακα μέσω της σχέσης $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\eta}}_k \hat{\Sigma}_{\mathbf{xx}}^{-1/2}$

Για να συγκρίνουν τη μέθοδο $SAVE$ με τη μέθοδο SIR , οι Cook and Weisberg (1991), έκαναν μια μελέτη προσομοίωσης χρησιμοποιώντας το μοντέλο $y = (\mu + 2^{1/2} \mathbf{z}_1 + 2^{1/2} \mathbf{z}_2)^2$ όπου τα \mathbf{z}_1 και \mathbf{z}_2 είναι διανύσματα με 120 ανεξάρτητες και ομοιόμορφα κατανεμημένες $N(0,1)$ μεταβλητές. Παρατηρούμε ότι για την τιμή $\mu = 0$, η y είναι τετραγωνική συνάρτηση ως προς τον γραμμικό συνδυασμό $\boldsymbol{\eta}^T \mathbf{z}$, όπου $\boldsymbol{\eta}^T = (1, 1)$ είναι το διάνυσμα που διατρέχει τον υποχώρο αποτελεσματικής μείωσης των διαστάσεων ενώ όταν το μ αυξάνει τότε η y τείνει να γίνει γραμμική συνάρτηση του $\boldsymbol{\eta}^T \mathbf{z}$. Η μελέτη προσομοίωσης έδειξε ότι η μέθοδος SIR απέτυχε στις περιπτώσεις όπου η τιμή του μ ήταν μικρή ενώ η μέθοδος $SAVE$ επέδειξε ικανοποιητικά αποτελέσματα για όλο το εύρος των τιμών του μ .

Οι Cook and Weisberg (1991), πρότειναν τον έλεγχο των μεταθέσεων για τον έλεγχο της σημαντικότητας τόσο για τη μέθοδο SIR όσο και για τη $SAVE$. Και για τις δυο μεθόδους τα slices καθορίζονται από τις τιμές της εξαρτημένης μεταβλητής y . Για να εκτιμήσουμε την κατανομή των μεταθέσεων για το μέσο όρο των $(p - K)$ μικρότερων ιδιοτιμών του πίνακα συνδιακυμάνσεων του πίνακα $\hat{\mathbf{V}}$, που όπως έχουμε ήδη αναφέρει

συμβολίζουμε με $\bar{\lambda}_{(p-K)}$, αντικαθιστούμε την y με μια τυχαία μετάθεση αυτής. Στη συνέχεια υπολογίζουμε την αντίστοιχη στατιστική συνάρτηση χρησιμοποιώντας είτε τη μέθοδο SIR είτε τη SAVE. Επαναλαμβάνουμε την παραπάνω διαδικασία έναν μεγάλο αριθμό φορών και καταγράφουμε το ποσοστό των φορών που η παρατηρούμενη τιμή της $\bar{\lambda}_{(p-K)}$ ξεπερνά την τιμή της $\bar{\lambda}_{(p-K)}$ που υπολογίστηκε από τα μετατιθέμενα δεδομένα. Η διαδικασία αυτή παρέχει μια εκτίμηση του p -value όταν τα ιδιοδιανύσματα ανήκουν στο χώρο αποτελεσματικής μείωσης των διαστάσεων. Για τη μέθοδο SIR η διαδικασία εκτιμά το p -value για οποιαδήποτε ελλειπτικά συμμετρική κατανομή και αν ακολουθεί το διάνυσμα x ενώ για τη μέθοδο SAVE μόνο αν το x ακολουθεί την πολυμεταβλητή κανονική κατανομή.

2.5 Principal Hessian directions

Εκτός από την τμηματική αντίστροφη παλινδρόμηση, ο Li (1992) πρότεινε και μια άλλη μέθοδο η οποία ονομάζεται *Principal Hessian directions* (pHd) για την εύρεση αντίστροφων δομών και η οποία παρέχει έναν έλεγχο για τη διάσταση του προβλήματος παλινδρόμησης. Η μέθοδος επεκτάθηκε από τον Cook (1998) και, όπως η μέθοδος SAVE, απαιτεί τόσο τη συνθήκη της γραμμικότητας της συνάρτησης παλινδρόμησης όσο και τη συνθήκη της σταθερής συνάρτησης διακύμανσης.

Έστω $\hat{\eta}_1, \dots, \hat{\eta}_p$ τα ιδιοδιανύσματα που αντιστοιχούν στα τετράγωνα των διατεταγμένων κατά φθίνουσα σειρά ιδιοτιμών του πίνακα

$$pHd = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}) \mathbf{z}_i \mathbf{z}_i^T,$$

όπου \mathbf{z} είναι το τυποποιημένο διάνυσμα. Το διάνυσμα των εκτιμημένων συντελεστών δίνεται από τη σχέση $\hat{\beta}_k = \hat{\eta}_k \hat{\Sigma}_{xx}^{-1/2}$, όπου $\hat{\Sigma}_{xx}$ είναι ο πίνακας δειγματικής διακύμανσης.

Η μέθοδος pHd ως μέθοδος μείωσης των διαστάσεων ενός προβλήματος παλινδρόμησης είναι καλύτερη όταν οι ανεξάρτητες μεταβλητές αντιστοιχούν σε μη γραμμικές τάσεις παρά όταν αντιστοιχούν σε γραμμικές. Για το λόγο αυτό η μέθοδος φαίνεται να είναι καλύτερη όταν χρησιμοποιείται ως διαγνωστικός έλεγχος για κάποιο μοντέλο, αντικαθιστώντας την

εξαρτημένη μεταβλητή y με τα κατάλοιπα e από ένα προσαρμοσμένο γραμμικό μοντέλο. Όταν χρησιμοποιείται κατά αυτόν τον τρόπο, ο αριθμός των γραμμικών συνδυασμών είναι ο μικρότερος αριθμός των επαρκών ανεξάρτητων μεταβλητών από την παλινδρόμηση των καταλοίπων e με το διάνυσμα των ανεξάρτητων μεταβλητών \mathbf{x} .

Ο έλεγχος που χρησιμοποιείται για τον προσδιορισμό των διαστάσεων της παλινδρόμησης, χρησιμοποιεί τη στατιστική συνάρτηση

$$n\Sigma_{\mathbf{xx}}^{-1/2} \sum_{j=m+1}^p \lambda_j^2,$$

η οποία ασυμπτωτικά ακολουθεί την χ^2 κατανομή με $(p-K)(p-K-1)/2$ βαθμούς ελευθερίας.

2.6 Συμπεράσματα

Στην παράγραφο αυτή, θα προσπαθήσουμε να κάνουμε μια σύντομη κριτική των παραπάνω μεθόδων. Αν και οι τρεις μέθοδοι δίνουν χρήσιμα πρακτικά αποτελέσματα, όλες έχουν κάποια μειονεκτήματα. Η μέθοδος *SIR*, πλεονεκτεί έναντι των υπολοίπων όταν υπάρχουν γραμμικές τάσεις στη συνάρτηση παλινδρόμησης $E(y|\mathbf{x})$. Ο δειγματικός συντελεστής συσχέτισης μεταξύ των προσαρμοσμένων με τη μέθοδο των ελαχίστων τετραγώνων τιμών και του πρώτου παράγοντα που δίνει η *SIR*, είναι τυπικά αρκετά μεγάλος. Η μέθοδος αυτή, επιπλέον, μπορεί να δώσει καλά αποτελέσματα στην περίπτωση που η συνάρτηση διακύμανσης $\text{var}(y|\mathbf{x})$ δεν είναι σταθερή. Παρόλα αυτά η μέθοδος *SIR* δεν είναι γενικά αποτελεσματική στην αναγνώριση της καμπυλότητας στη συνάρτηση παλινδρόμησης των καταλοίπων $E(e|\mathbf{x})$.

Η μέθοδος *SAVE*, βρίσκει πάντοτε εκτός από τις επαρκείς μεταβλητές που βρίσκουν οι μέθοδοι *SIR* και *rHd*, και όποιες άλλες υπάρχουν στην παλινδρόμηση. Παρόλα αυτά το στοιχείο αυτό της μεθόδου *SAVE*, έχει και κάποιο κόστος. Συγκεκριμένα η *SAVE*, ψάχνει σε μια μεγαλύτερη κλάση γραμμικών συνδυασμών που περιλαμβάνει τη μικρότερη κλάση στην οποία ψάχνουν *SIR* και *rHd*. Πολλές φορές η *SAVE*, δυσκολεύεται να βρει σχετικά

ευθείες δομές τις οποίες οι άλλες δυο μέθοδοι βρίσκουν εύκολα, συνήθως όταν ο αριθμός των ανεξάρτητων μεταβλητών είναι μεγάλος.

Για τη μέθοδο pHd , μπορεί γενικά να λεχθεί ότι είναι μια μέθοδος που ειδικεύεται στην αναγνώριση της καμπυλότητας στη συνάρτηση παλινδρόμησης των καταλοίπων $E(e | \mathbf{x})$. Δεν φαίνεται όμως να δουλεύει τόσο καλά όσο οι μέθοδοι *SIR* και *SAVE*, για την εύρεση των κατευθύνσεων όταν η συνάρτηση διακύμανσης των καταλοίπων $\text{var}(e | \mathbf{x})$ δεν είναι σταθερή.



Κεφάλαιο 3:

Εφαρμογές

3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιαστούν δυο πρακτικές εφαρμογές, έτσι ώστε να γίνει κατανοητό το πώς χρησιμοποιούνται στην πράξη τα όσα έχουμε αναφέρει στα δυο προηγούμενα κεφάλαια. Η δομή που θα ακολουθήσουμε θα είναι η εξής: Στην αρχή θα δίνεται μια συνοπτική περιγραφή του προβλήματος. Στη συνεχεία θα παρουσιάζεται η σχέση μεταξύ της εξαρτημένης μεταβλητής με τις ανεξάρτητες μεταβλητές μέσω κατάλληλων διαγραμμάτων ενώ στο τέλος θα εφαρμοσθούν τόσο γραφικές μέθοδοι όσο και η μέθοδος της τμηματικής αντίστροφης παλινδρόμησης (*SIR*) για να βρούμε τη δομική διάσταση κάθε προβλήματος.

3.2 Ανάλυση προβλήματος Smoking and Cancer

3.2.1 Περιγραφή του προβλήματος

Το πρώτο πρόβλημα, με το οποίο θα ασχοληθούμε, αφορά τον κατά κεφαλήν αριθμό των τσιγάρων που καπνίσθηκαν (πωλήθηκαν) σε 43 πολιτείες της Αμερικής και στην περιφέρεια της Κολούμπια το 1960 μαζί με τον αριθμό των θανάτων ανά χίλιους κατοίκους από διάφορες μορφές καρκίνου. Τα δεδομένα αυτά βρίσκονται στο διαδίκτυο, στη διεύθυνση <http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html>. Συγκεκριμένα, στο αρχείο που θα χρησιμοποιήσουμε περιέχονται 6 μεταβλητές, των οποίων η περιγραφή δίνεται στον Πίνακα 3.1. Στην ανάλυση που θα ακολουθήσει, θα χρησιμοποιήσουμε ως εξαρτημένη μεταβλητή τη *Cig* που δηλώνει τον αριθμός καπνισμένων τσιγάρων ανά άτομο σε χιλιάδες. Ως ανεξάρτητες μεταβλητές θα χρησιμοποιήσουμε μόνο τις μεταβλητές *Blad*, *Lung*, *Kid* και *Leuk*.

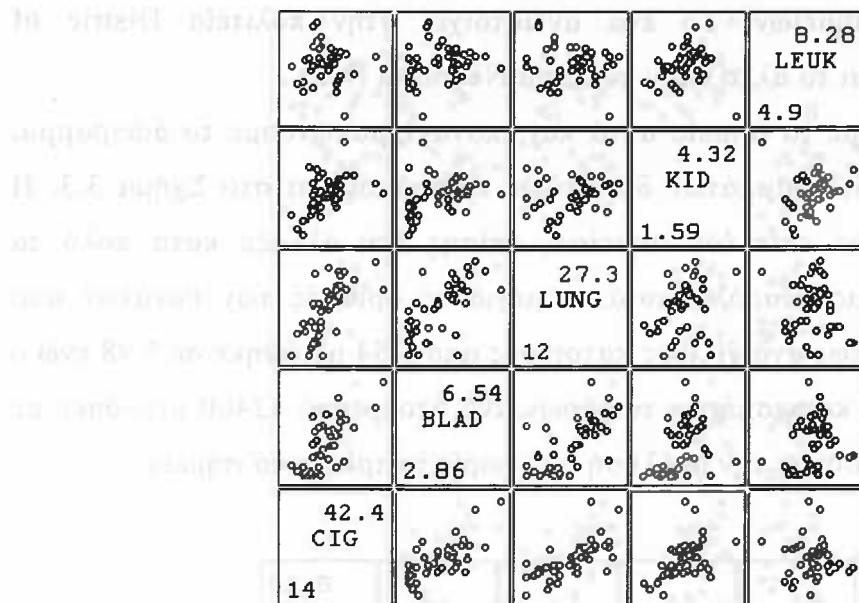
3.2.2 Διδιάστατη γραφική απεικόνιση των δεδομένων

Θα ξεκινήσουμε την ανάλυσή μας, κατασκευάζοντας τον πίνακα διαγραμμάτων διασποράς, ο οποίος όπως έχουμε ήδη αναφέρει αποτελείται από μια σειρά διδιάστατων διαγραμμάτων διασποράς. Ο πίνακας διαγραμμάτων διασποράς εμφανίζεται στο Σχήμα 3.1. Παρατηρούμε ότι εκτός από τη διαγώνιο, κάθε πλαίσιο περιέχει ένα διάγραμμα διασποράς. Οι τιμές που εμφανίζονται μαζί με το όνομα κάθε μεταβλητής στη διαγώνιο του πίνακα, είναι η ελάχιστη και η μέγιστη τιμή κάθε μεταβλητής. Για παράδειγμα, ο αριθμός των θανάτων από καρκίνο του πνεύμονα κυμαίνεται από 12 το ελάχιστο έως 27.3 το μέγιστο ανά χιλιούς κατοίκους.

Όνομα μεταβλητής	Τύπος	Μέγεθος	Περιγραφή
State	Ποιοτική	44	Πολιτεία
Cig	Αριθμητική	44	Αριθμός καπνισμένων τσιγάρων ανά άτομο σε χιλιάδες
Blad	Αριθμητική	44	Θάνατοι ανά 1000 κατοίκους από καρκίνο της κύστεως
Lung	Αριθμητική	44	Θάνατοι ανά 1000 κατοίκους από καρκίνο του πνεύμονα
Kid	Αριθμητική	44	Θάνατοι ανά 1000 κατοίκους από καρκίνο νεφρών
Leuk	Αριθμητική	44	Θάνατοι ανά 1000 κατοίκους από λευχαιμία

Πίνακας 3.1: Περιγραφή μεταβλητών προβλήματος *Smoking and Cancer*

Σημειώνουμε ότι τα διαγράμματα που βρίσκονται πάνω από την κύρια διαγώνιο του πίνακα, αποτελούν τις αντίστροφες απεικονίσεις (mirror images) των διαγραμμάτων που βρίσκονται κάτω από την κύρια διαγώνιο. Για παράδειγμα, τα γραφήματα που βρίσκονται στην πρώτη στήλη του πίνακα είναι οι αντίστροφες απεικονίσεις των γραφημάτων της τελευταίας γραμμής του πίνακα. Για παράδειγμα, το κάτω δεξιά διάγραμμα {Leuk, Cig} είναι η αντίστροφη απεικόνιση του πάνω αριστερά διαγράμματος {Cig, Leuk}.



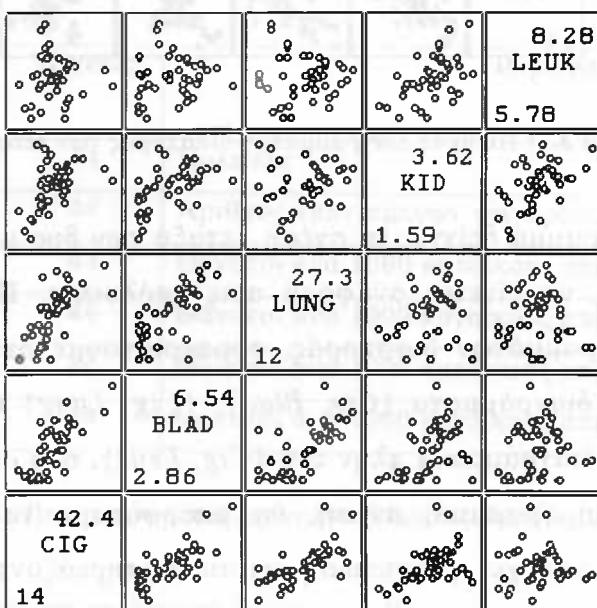
Σχήμα 3.1: Πίνακας διαγραμμάτων διασποράς των δεδομένων

Κάθε διάγραμμα δείχνει τη σχέση μεταξύ των δυο μεταβλητών που το αποτελούν χωρίς να γίνεται αναφορά στις υπόλοιπες. Παρατηρώντας τον πίνακα των διαγραμμάτων διασποράς, συμπεραίνουμε ότι γραμμική σχέση απεικονίζουν τα διαγράμματα $\{Cig, Blad\}$, $\{Cig, Lung\}$ και $\{Blad, Lung\}$. Για τα υπόλοιπα διαγράμματα πλην του $\{Cig, Leuk\}$, στο οποίο παρατηρούμε ξεκάθαρα μια μη γραμμική σχέση, θα μπορούσαμε να πούμε ότι είναι γραμμικά αν και υπάρχει η εντύπωση ότι αυτά επηρεάζονται σημαντικά από κάποιο ή κάποια απομονωμένα σημεία.

Εξετάζοντας τα διαγράμματα της πρώτης γραμμής του πίνακα των διαγραμμάτων διασποράς, παρατηρούμε ένα σημείο να βρίσκεται μακριά από τα υπόλοιπα. Εάν διαγράψουμε αυτό το σημείο, που αντιστοιχεί στην πολιτεία Alaska (AK), και επανακλιμακώσουμε το διάγραμμα, παίρνουμε τον πίνακα διαγραμμάτων διασποράς του Σχήματος 3.2. Παρατηρούμε ότι η διαγραφή του σημείου, δεν άλλαξε κατά πολύ τα συμπεράσματά μας. Μια σημαντική διαφορά είναι ότι ο μέγιστος αριθμός των θανάτων από καρκίνο των νεφρών ανά χίλιους κατοίκους από 4.32 μειώθηκε σε 3.62 ενώ ο ελάχιστος αριθμός των θανάτων από λευχαιμία ανά χίλιους κατοίκους από 4.9 αυξήθηκε σε 5.78. Εξετάζοντας την τελευταία γραμμή του νέου πίνακα διαγραμμάτων διασποράς, παρατηρούμε επίσης, την ύπαρξη δυο επιπλέον

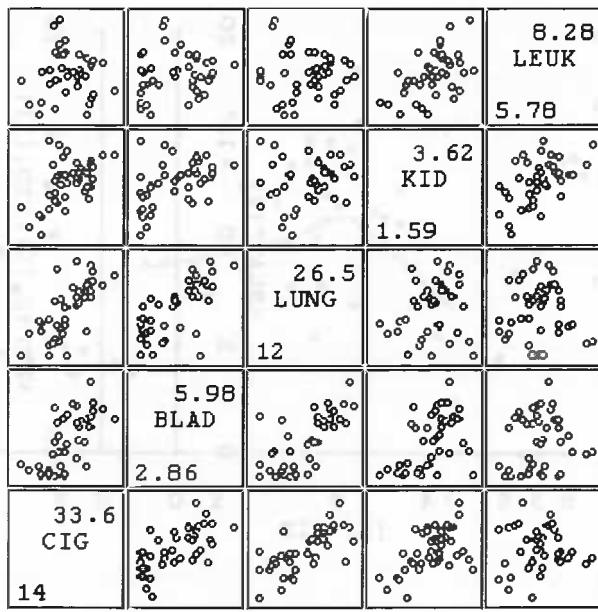
απομονωμένων σημείων. Το ένα αντιστοιχεί στην πολιτεία District of Columbia (DC) και το άλλο στην πολιτεία Nebraska (NE).

Διαγράφουμε τα σημεία αυτά και επανακλιμακώνουμε το διάγραμμα. Ο νέος πίνακας διαγραμμάτων διασποράς παρουσιάζεται στο Σχήμα 3.3. Η διαγραφή των δυο επιπλέον σημείων, επίσης δεν άλλαξε κατά πολύ τα συμπεράσματά μας. Παρόλα αυτά, ο μέγιστος αριθμός των θανάτων από καρκίνο της κύστεως ανά χίλιους κατοίκους από 6.54 μειώθηκε σε 5.98 ενώ ο μέγιστος αριθμός καπνισμένων τσιγάρων ανά άτομο από 42400 μειώθηκε σε 33600. Θα συνεχίσουμε την ανάλυσή μας χωρίς τα τρία αυτά σημεία.



Σχήμα 3.2: Πίνακας διαγραμμάτων διασποράς μετά τη διαγραφή του απομονωμένου σημείου

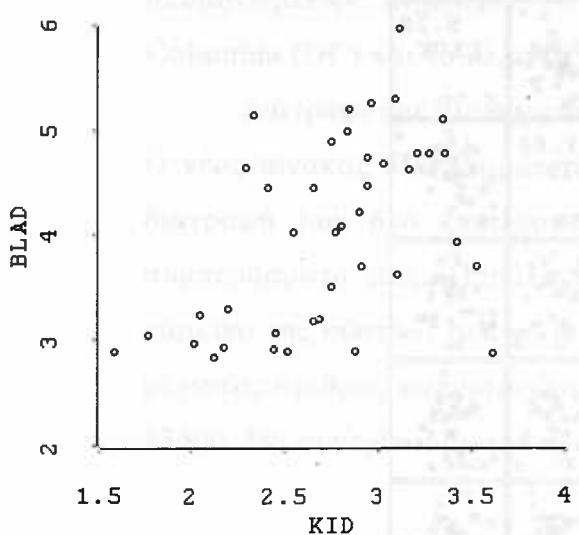
Για να είναι αποτελεσματικές οι γραφικές μέθοδοι, θα πρέπει όλες οι συναρτήσεις παλινδρόμησης της μορφής $E(x_j | x_k)$ για όλα τα j και k να είναι γραμμικές συναρτήσεις των x_k . Αν ισχύει αυτό, τότε όλα τα διαγράμματα της μορφής $\{x_k, x_j\}$ στον πίνακα διαγραμμάτων διασποράς θα δείχνουν γραμμική σχέση. Συνεπώς, όταν δεν εμφανίζεται γραμμική σχέση, θα πρέπει οι ανεξάρτητες μεταβλητές να μετασχηματίζονται έτσι ώστε να επιτυγχάνεται όσο το δυνατόν καλύτερα η γραμμικότητα.



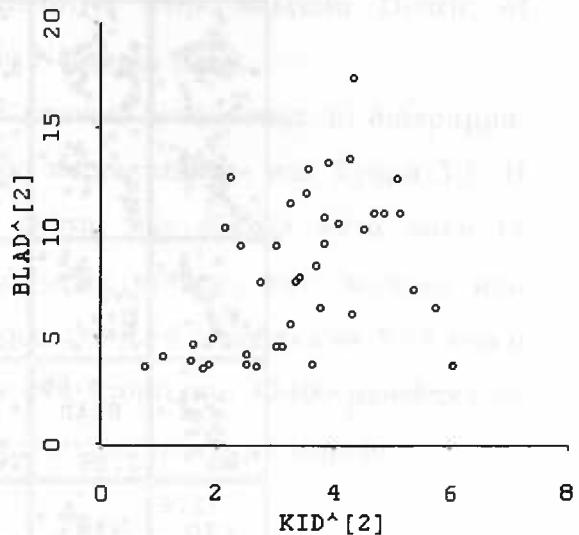
Σχήμα 3.3: Πίνακας διαγραμμάτων διασποράς μετά τη διαγραφή
των δυο επιπλέον απομονωμένων σημείων

Μια λύση είναι να χρησιμοποιήσουμε τους μετασχηματισμούς δύναμης, καθώς όλα τα διαγράμματα του πίνακα φαίνεται να είναι μονότονα. Θεωρούμε αρχικά το διάγραμμα $\{Kid, Blad\}$ από το οποίο παρατηρούμε ότι η οι παρατηρήσεις είναι διασκορπισμένες. Άρα για να συμπυκνώσουμε τις μεγάλες τιμές, θα πρέπει να χρησιμοποιήσουμε μεγάλη τιμή για την παράμετρο λ . Το πρόγραμμα Arc , μας δίνει τη δυνατότητα να μετασχηματίσουμε τη μεταβλητή και το αποτέλεσμα του μετασχηματισμού φαίνεται σε όλα τα διαγράμματα που περιέχουν τη συγκεκριμένη μεταβλητή.

Στο Σχήμα 3.4 εμφανίζεται το διάγραμμα διασποράς των αρχικών μεταβλητών Kid και $Blad$, ενώ στο Σχήμα 3.5 παρουσιάζεται το αντίστοιχο διάγραμμα μετά το μετασχηματισμό των δυο μεταβλητών χρησιμοποιώντας την τιμή $\lambda = 2$. Έχουμε δηλαδή υψώσει τις τιμές των δυο μεταβλητών στην τιμή 2. Παρατηρώντας το Σχήμα 3.5, παρατηρούμε ότι αν και τα πράγματα έχουν βελτιωθεί μάλλον δεν έχουμε πετύχει τη γραμμικότητα.

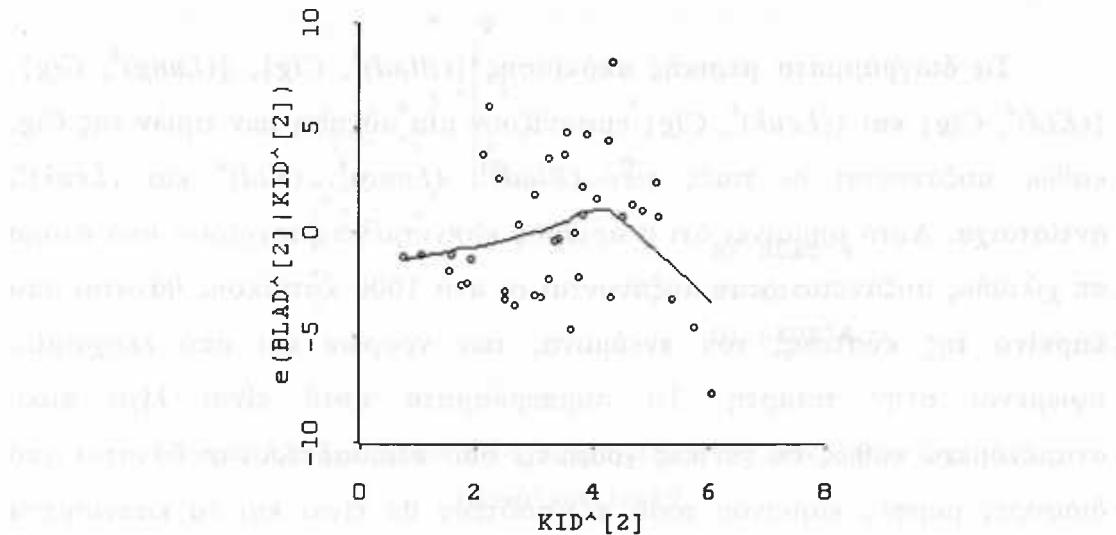


Σχήμα 3.4: Διάγραμμα διασποράς των μεταβλητών
Kid και *Blad*



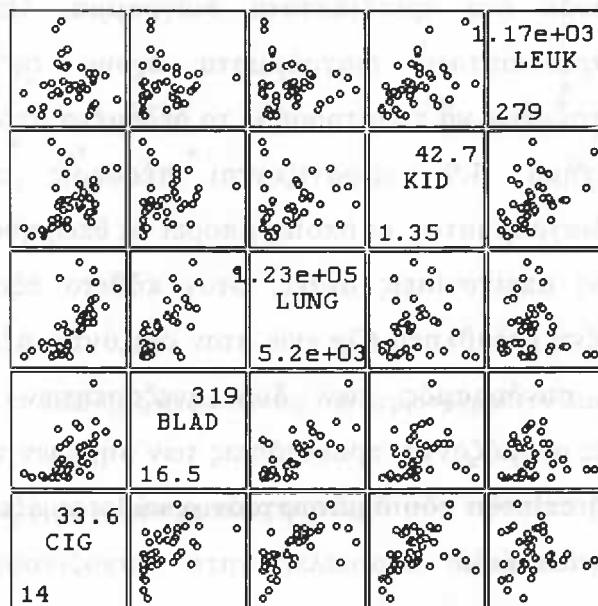
Σχήμα 3.5: Διάγραμμα διασποράς των μετασχηματισμένων
μεταβλητών *Kid* και *Blad*

Ένας τρόπος για να ελέγξουμε αν όντως έχουμε πετύχει τη γραμμικότητα, είναι να κατασκευάσουμε το διάγραμμα καταλοίπων. Για να δημιουργήσουμε το διάγραμμα καταλοίπων, αφαιρούμε από το διάγραμμα του Σχήματος 3.5 τη γραμμική τάση (Rem Lin Trend), οπότε οι τιμές του κάθετου άξονα αντικαθίστανται από τα κατάλοιπα της παλινδρόμησης με εξαρτημένη μεταβλητή τη μετασχηματισμένη μεταβλητή *Kid* και ανεξάρτητη μεταβλητή τη μετασχηματισμένη *Kid*, δηλαδή τη $(Kid)^2$. Στη συνέχεια προσαρμόζουμε στο διάγραμμα καταλοίπων μια ομαλή καμπύλη *lowess*. Από το Σχήμα 3.6, όπου εμφανίζεται το διάγραμμα καταλοίπων με προσαρμοσμένη μια *lowess* με παράμετρο $f = 0.8$ καμπύλη, συμπεραίνουμε ότι ο μετασχηματισμός που χρησιμοποιήσαμε δεν πέτυχε τη γραμμικότητα. Συνεπώς θα πρέπει να χρησιμοποιήσουμε μια ακόμη μεγαλύτερη τιμή για το λ .



Σχήμα 3.6: Διάγραμμα καταλοίπων με προσαρμοσμένη *lowess* καμπύλη

Συνεχίζουμε την ανάλυση, μετασχηματίζοντας όλες τις ανεξάρτητες μεταβλητές χρησιμοποιώντας την τιμή $\lambda = 4$. Αυτό σημαίνει ότι αντικαθιστούμε τις τιμές των μεταβλητών με την τιμή που προκύπτει εάν υψώσουμε τις τιμές αυτές στην τετάρτη. Προσεγγιστικά μπορούμε να πούμε, ότι σε όλα τα διαγράμματα διασποράς, εμφανίζεται μια γραμμική σχέση.

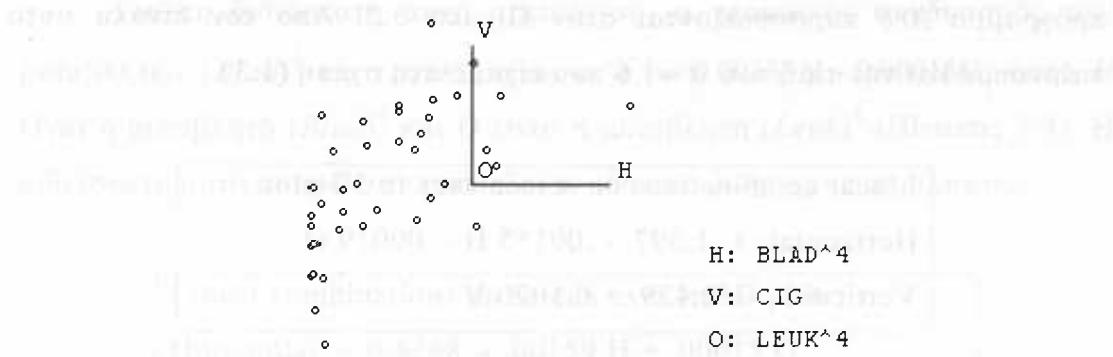


Σχήμα 3.7: Πίνακας διαγραμμάτων διασποράς με μετασχηματισμένες ανεξάρτητες μεταβλητές

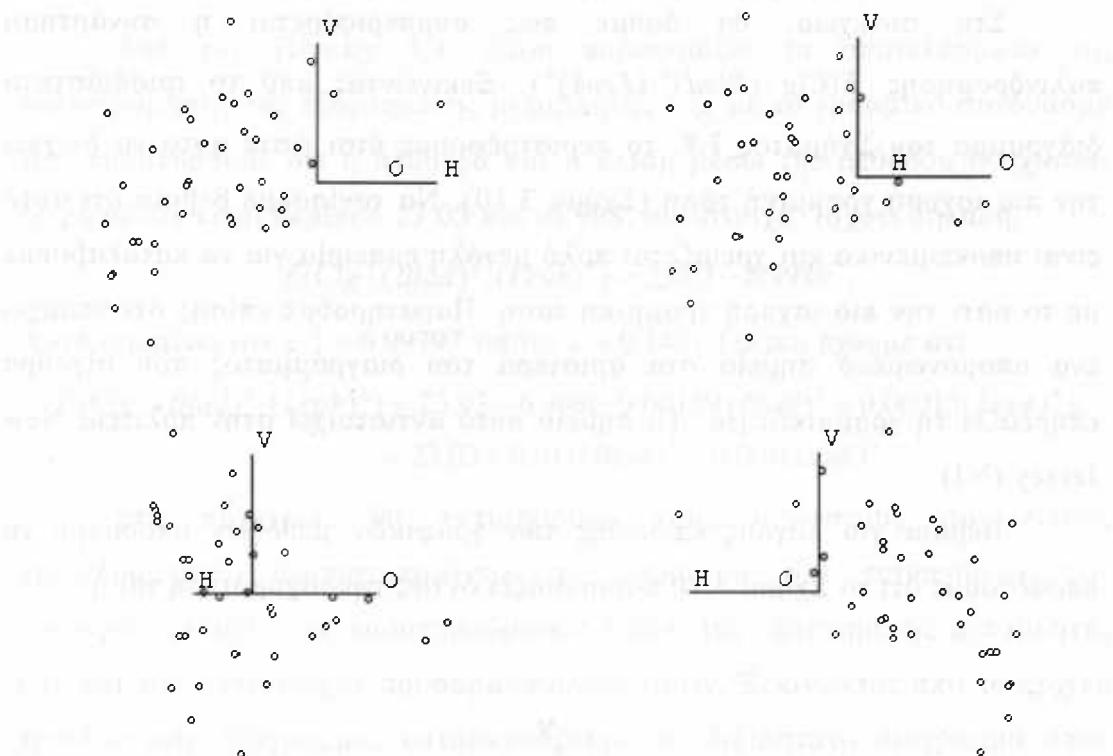
Τα διαγράμματα μερικής απόκρισης $\{(Blad)^4, Cig\}$, $\{(Lung)^4, Cig\}$, $\{(Kid)^4, Cig\}$ και $\{(Leuk)^4, Cig\}$ εμφανίζουν μια αύξηση των τιμών της *Cig*, καθώς αυξάνονται οι τιμές των $(Blad)^4$, $(Lung)^4$, $(Kid)^4$ και $(Leuk)^4$, αντίστοιχα. Αυτό σημαίνει ότι ο αριθμός καπνισμένων τσιγάρων ανά άτομο σε χιλιάδες αυξάνεται όταν αυξάνονται οι ανά 1000 κατοίκους θάνατοι από καρκίνο της κύστεως, του πνεύμονα, των νεφρών και από λευχαιμία, υψωμένοι στην τετάρτη. Τα συμπεράσματα αυτά είναι λίγο πολύ αναμενόμενα καθώς σε γενικές γραμμές, όσο περισσότεροι οι θάνατοι από διάφορες μορφές καρκίνου τόσο περισσότερα θα είναι και τα καπνισμένα τσιγάρα, αφού είναι αποδεδειγμένο ότι το κάπνισμα σχετίζεται με διάφορες μορφές καρκίνου.

3.2.3 Τρισδιάστατη γραφική απεικόνιση

Στο Σχήμα 3.8 παρουσιάζουμε ένα τρισδιάστατο διάγραμμα, όπου στον κάθετο άξονα (*V*) απεικονίζεται η εξαρτημένη μεταβλητή *Cig*, στον οριζόντιο άξονα (*H*) η μεταβλητή $(Blad)^4$ ενώ στον εκτός της οθόνης άξονα (*O*) απεικονίζεται η μεταβλητή $(Leuk)^4$. Σημειώνουμε ότι θα μπορούσαμε να χρησιμοποιήσουμε οποιεσδήποτε από τις μετασχηματισμένες μεταβλητές για να κατασκευάσουμε ένα τρισδιάστατο διάγραμμα. Όπως ήδη έχουμε αναφέρει, τα τρισδιάστατα διαγράμματα έχουν τη δυνατότητα να περιστρέφονται έτσι ώστε να παρατηρούμε τα δεδομένα από διάφορες οπτικές γωνίες. Στο Σχήμα 3.9, εμφανίζονται τέσσερις απεικονίσεις του περιστραμμένου διαγράμματος, οι οποίες μπορεί να θεωρηθούν ως διδιάστατα διαγράμματα. Στις απεικονίσεις αυτές, στον κάθετο άξονα απεικονίζεται ακόμα η εξαρτημένη μεταβλητή *Cig* ενώ στον οριζόντιο άξονα απεικονίζεται ένας γραμμικός συνδυασμός των δυο ανεξάρτητων μεταβλητών. Οι απεικονίσεις αυτές ονομάζονται προεκτάσεις των σημείων του τρισδιάστατου διαγράμματος στο επίπεδο που δημιουργούν ο κάθετος άξονας και ο άξονας του γραμμικού συνδυασμού.



Σχήμα 3.8: Τρισδιάστατο διάγραμμα μεταξύ των μεταβλητών Cig , $(Blad)^4$ και $(Leuk)^4$



Σχήμα 3.9: Διάφορες απεικονίσεις του περιστραμμένου διαγράμματος

Για παράδειγμα, ο γραμμικός συνδυασμός των μεταβλητών $(Blad)^4$ και $(Leuk)^4$ που εμφανίζονται στην τελευταία διδιάστατη απεικόνιση του Σχήματος 3.8, είναι ο $h \approx -0.00155H - 0.00019O$, όπου H είναι η μεταβλητή $(Blad)^4$ και O είναι η μεταβλητή $(Leuk)^4$. Τα αποτελέσματα που δίνει το

πρόγραμμα *Arc* παρουσιάζονται στον Πίνακα 3.2. Από τον πίνακα αυτό παίρνουμε και την τιμή του $d \approx 1.6$ που είχαμε στη σχέση (1.3).

Linear combinations on screen axes in 3D plot.

Horizontal: + 1.597 - .00155 H - .00019 O

Vertical: - 2.429 + 0.1020 V

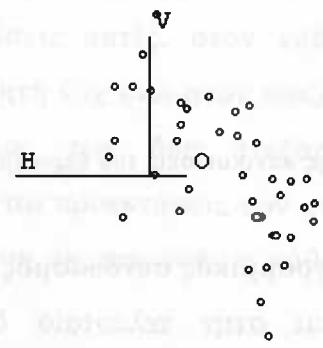
Πίνακας 3.2: Οι γραμμικοί συνδυασμού που αποτελούν

τους άξονες του διαγράμματος

3.2.4 Απεικόνιση της γραμμικής παλινδρόμησης

Στη συνέχεια, θα δούμε πως συμπεριφέρεται η συνάρτηση παλινδρόμησης $E(Cig | (Blad)^4, (Leuk)^4)$. Ξεκινώντας από το τρισδιάστατο διάγραμμα του Σχήματος 3.8, το περιστρέφουμε έτσι ώστε αυτό να δείχνει την πιο ισχυρή γραμμική τάση (Σχήμα 3.10). Να τονίσουμε βέβαια ότι αυτό είναι υποκειμενικό και χρειάζεται πολύ μεγάλη εμπειρία για να καταλάβουμε με το μάτι την πιο ισχυρή γραμμική τάση. Παρατηρούμε επίσης ότι υπάρχει ένα απομονωμένο σημείο στα αριστερά του διαγράμματος που σίγουρα επηρεάζει τη γραμμικότητα. Το σημείο αυτό αντιστοιχεί στην πολιτεία New Jersey (NJ).

Βέβαια για λόγους επίδειξης των γραφικών μεθόδων μπορούμε να υποθέσουμε ότι το Σχήμα 3.10 παρουσιάζει όντως την ισχυρότερη τάση.



Σχήμα 3.10: Η διδιάστατη απεικόνιση με την πιο ισχυρή γραμμική τάση

Για τη διδιάστατη αυτή απεικόνιση, ο γραμμικός συνδυασμός των μεταβλητών $(Blad)^4$ και $(Leuk)^4$ είναι ο $h^* \approx -0.00159H + 0.00015O$, όπου H είναι η μεταβλητή $(Blad)^4$ και O είναι η μεταβλητή $(Leuk)^4$ (Πίνακας 3.3). Η διδιάστατη αυτή απεικόνιση αποτελεί το ιδανικό διάγραμμα περίληψης.

Linear combinations on screen axes in 3D plot.

Horizontal: $+ 0.6268 - .00159 H + .00015 O$

Vertical: $- 2.429 + 0.1020 V$

Πίνακας 3.3: Οι γραμμικοί συνδυασμοί που αποτελούν τους
άξονες του διαγράμματος του Σχήματος 3.8

Από τον Πίνακα 3.4, όπου παρουσιάζει τα αποτελέσματα της παλινδρόμησης της εξαρτημένης μεταβλητής Cig με το γραμμικό συνδυασμό h^* , παρατηρούμε ότι η σταθερά και η κλίση μέσω της μεθόδου ελαχίστων τετραγώνων είναι περίπου 23.03 και -6.998, αντίστοιχα. Ισχύει δηλαδή,

$$E(Cig | (Blad)^4, (Leuk)^4) = 23.03 - 6.998h^*.$$

Αυτό σημαίνει ότι $c^{-1} = 6.99797$ οπότε $c = 0.143$. Τελικά έχουμε ότι

$$\begin{aligned} E(Cig | (Blad)^4, (Leuk)^4) &= 23.03 - 6.998(-0.00159(Blad)^4 + 0.00015(Leuk)^4) \\ &= 23.03 + 0.011(Blad)^4 - 0.001(Leuk)^4. \end{aligned}$$

Στη συνέχεια, θα εκτιμήσουμε τους άγνωστους συντελεστές παλινδρόμησης, ελαχιστοποιώντας το άθροισμα των τετραγώνων των διαφορών μεταξύ των παρατηρούμενων τιμών της εξαρτημένης μεταβλητής Cig και των αντίστοιχων προσαρμοσμένων τιμών. Ξεκινώντας από το αρχικό τρισδιάστατο διάγραμμα, κατασκευάζουμε το διδιάστατο διάγραμμα όπου στον κάθετο άξονα απεικονίζεται η εξαρτημένη μεταβλητή Cig και στον οριζόντιο άξονα το γραμμικό συνδυασμό h_{ols} που παράγεται από τη μέθοδο των ελαχίστων τετραγώνων.

Data set = smokingandcancer

Normal Regression

Kernel mean function = Identity

Response = CIG

Terms = (h^*)

Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
Constant	23.0316	0.639646	36.007	0.0000
h^*	-6.99797	1.00616	-6.955	0.0000

R Squared: 0.535263

Sigma hat: 3.84437

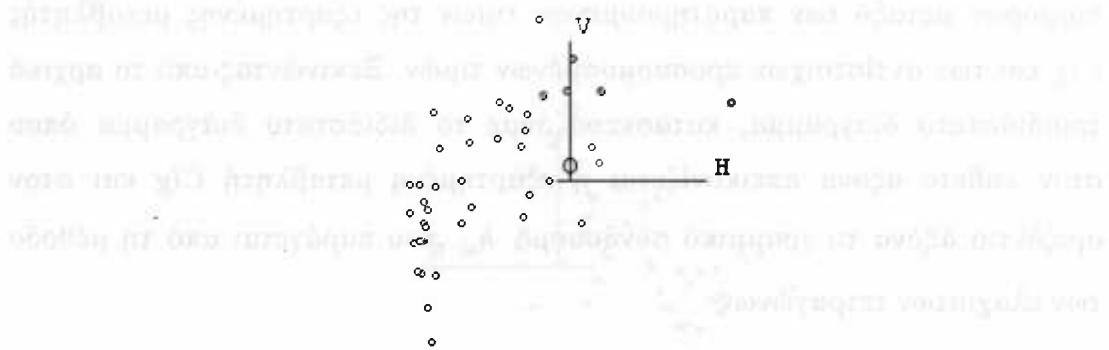
Number of cases: 44

Degrees of freedom: 42

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	1	714.921	714.921	48.37	0.0000
Residual	42	620.724	14.7791		

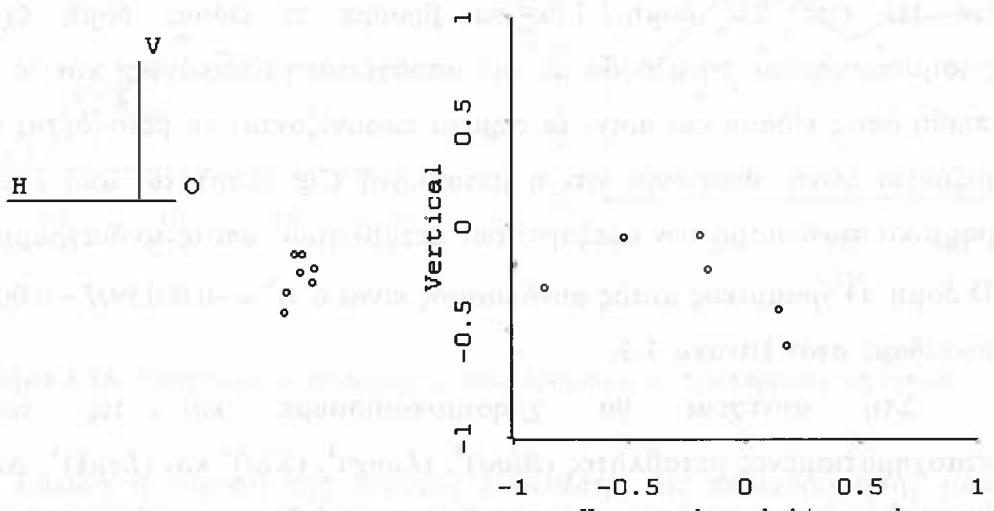
Πίνακας 3.4: Αποτελέσματα παλινδρόμησης της Cig με την h^*



Σχήμα 3.11: Διδιάστατη απεικόνιση $\{Cig, h_{ols}\}$

Συγκρίνοντας τα Σχήματα 3.11 και 3.10, παρατηρούμε ότι το ένα είναι περίπου η αντίστροφη απεικόνιση του άλλου. Δηλαδή εάν κοιτάξουμε το Σχήμα 3.10 από την πίσω του μεριά, θα δούμε περίπου τη διδιάστατη απεικόνιση που παρουσιάζεται στο Σχήμα 3.11. Αυτό σημαίνει ότι η γραμμική τάση που επιλέξαμε με το μάτι στο Σχήμα 3.10, ήταν όντως από τις ισχυρότερες αν όχι η πιο ισχυρή.

Για να ελέγξουμε κατά πόσο το διάγραμμα περίληψης του Σχήματος 3.10 είναι ικανοποιητικό, ακολουθούμε τη διαδικασία της τμηματοποίησης που έχουμε περιγράψει στο Κεφάλαιο 1. Ξεκινώντας από το Σχήμα 3.10, κατασκευάζουμε την ασυχέτιστη απεικόνιση του τρισδιάστατου διαγράμματος και δημιουργούμε ένα slicer με ποσοστό 0.2. Στο Σχήμα 3.12 απεικονίζουμε ένα slice στο διάγραμμα περίληψης και τις αντίστοιχες παρατηρήσεις στην ασυχέτιστη απεικόνιση του διαγράμματος περίληψης. Από τη συγκεκριμένη ασυχέτιστη απεικόνιση, επειδή τα σημεία εμφανίζονται να βρίσκονται σε μια οριζόντια ζώνη, θεωρούμε ότι το διάγραμμα περίληψης είναι επαρκές. Να τονίσουμε βέβαια ότι επειδή τα αρχικά μας δεδομένα σε κάποια slices είναι λίγα, η ασυχέτιστη απεικόνιση μπορεί να οδηγήσει σε λάθος συμπεράσματα και συνεπώς σε λανθασμένη απόρριψη του διαγράμματος περίληψης.



(a) Διάγραμμα περίληψης

(b) Ασυχέτιστη απεικόνιση

Σχήμα 3.12: Διάγραμμα περίληψης και ασυχέτιστη απεικόνιση αυτού

3.2.5. Εύρεση της δομής του προβλήματος

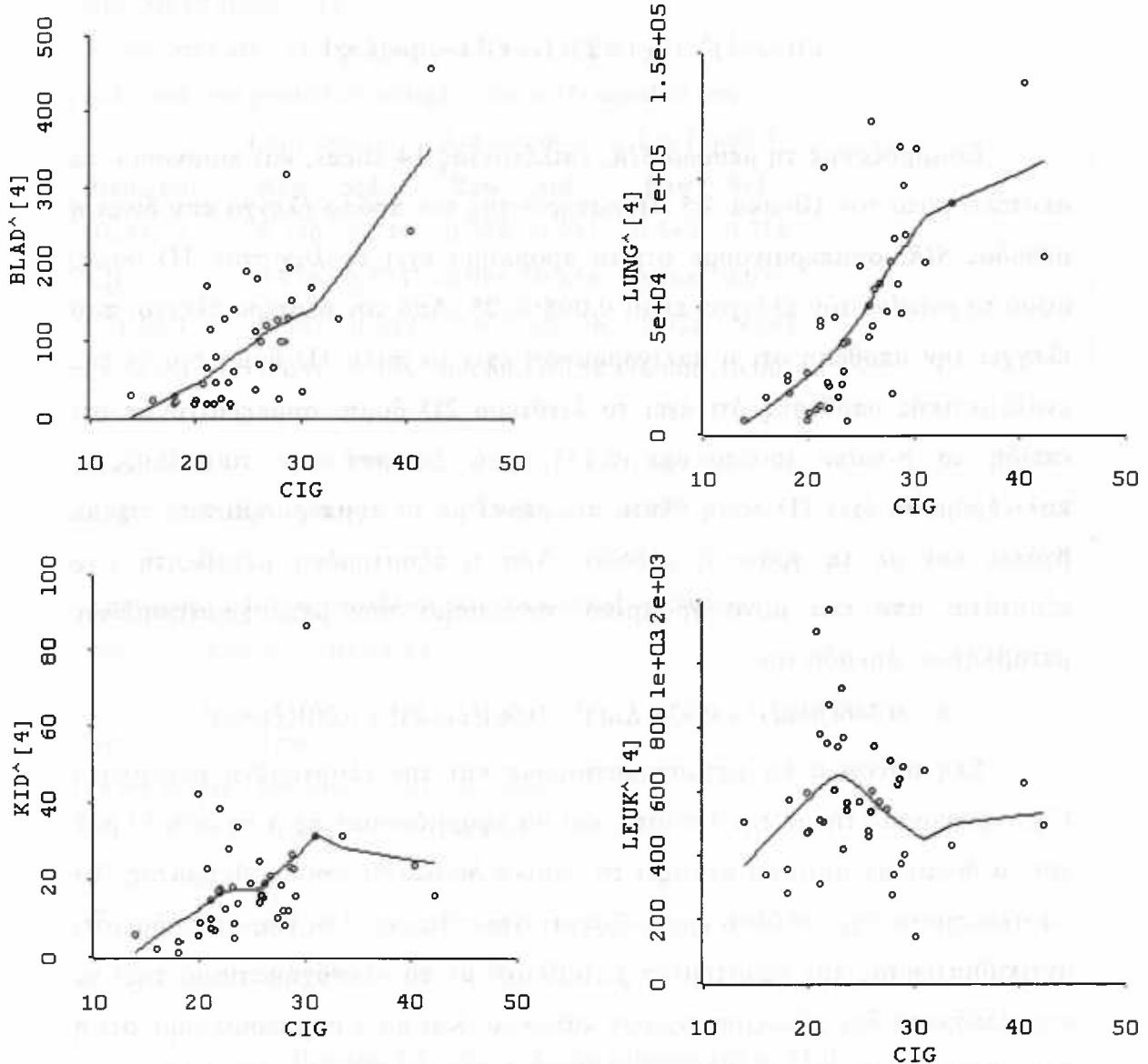
Στα προηγούμενα, είδαμε πώς μπορούμε να χρησιμοποιήσουμε τα διαγράμματα στην περίπτωση που θεωρούμε ότι τα δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Στην παράγραφο αυτή, θα προσπαθήσουμε τόσο γραφικά όσο και εφαρμόζοντας τη μέθοδο *SIR*, να βρούμε τη δομική διάσταση του προβλήματος παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής *Cig* και των μετασχηματισμένων μεταβλητών (*Blad*)⁴, (*Lung*)⁴, (*Kid*)⁴ και (*Leuk*)⁴, χωρίς να υποθέτουμε την ισχύ κάποιου συγκεκριμένου μοντέλου. Να θυμίσουμε ότι το πρόβλημα αυτό, επειδή έχει τέσσερις ανεξάρτητες μεταβλητές μπορεί να έχει μέχρι 4D δομή.

Περιστρέφοντας και παρατηρώντας το τρισδιάστατο διάγραμμα του Σχήματος 3.8, δηλαδή το διάγραμμα $\{(Blad)^4, Cig, (Leuk)^4\}$ συμπεραίνουμε ότι αυτό παρουσιάζει κάποιο συστηματικό πρότυπο. Συγκεκριμένα φαίνεται ότι σχεδόν όλες οι παρατηρήσεις είναι συγκεντρωμένες κάτω δεξιά ενώ κάποιες απομονωμένες τιμές βρίσκονται πάνω αριστερά. Έχουμε ήδη αναφερθεί στην παρατήρηση που αντιστοιχεί στην πολιτεία New Jersey (NJ).

Αυτό, αν και μπορεί να οφείλεται στο ότι ο κύριος όγκος των παρατηρήσεων βρίσκεται κάτω δεξιά, σημαίνει ότι έχουμε αυξανόμενη διακύμανση. Αυτό οδηγεί στο συμπέρασμα ότι πρέπει να απορρίψουμε την υπόθεση της μηδενικής δομικής διάστασης. Συνεπώς το διάγραμμα θα έχει είτε 1D είτε 2D δομή. Για να βρούμε τι είδους δομή έχει θα χρησιμοποιήσουμε τη μέθοδο με την ασυσχέτιστη απεικόνιση και το *slicer*. Επειδή όπως είδαμε και πριν, τα σημεία εμφανίζονται να βρίσκονται σε μια οριζόντια ζώνη, θεωρούμε ότι η μεταβλητή *Cig* εξαρτάται από ένα μόνο γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών, οπότε το διάγραμμα έχει 1D δομή. Ο γραμμικός αυτός συνδυασμός είναι ο $h^* \approx -0.00159H + 0.00015O$, που είδαμε στον Πίνακα 3.3.

Στη συνέχεια θα χρησιμοποιήσουμε και τις τέσσερις μετασχηματισμένες μεταβλητές (*Blad*)⁴, (*Lung*)⁴, (*Kid*)⁴ και (*Leuk*)⁴. Από τον πίνακα διαγραμμάτων διασποράς του Σχήματος 3.7, συμπεραίνουμε ότι εκτός από μερικά απομακρυσμένα σημεία ισχύει η υπόθεση των γραμμικών ανεξάρτητων μεταβλητών. Από την πρώτη στήλη του πίνακα, που παρουσιάζει τα διαγράμματα αντίστροφης παλινδρόμησης, συμπεραίνουμε ότι επειδή και τα τέσσερα διαγράμματα δείχνουν περίπου το ίδιο πρότυπο, τα

πρόβλημα πρέπει να έχει 1D δομή. Τα τέσσερα διαγράμματα μερικής απόκρισης, δίνονται στο Σχήμα 3.13 ενώ σε κάθε διάγραμμα έχουμε προσαρμόσει μια ομαλή καμπύλη *lowess* με παράμετρο 0.7.



Σχήμα 3.13: Διαγράμματα αντίστροφης παλινδρόμησης με προσαρμοσμένη *lowess*

Επειδή η εύρεση της δομικής διάστασης της παλινδρόμησης μέσω διαγραμμάτων βασίζεται πολύ σε οπτικές απεικονίσεις, θα εφαρμόσουμε και τη μέθοδο της τυμηματικής αντίστροφης παλινδρόμησης (*SIR*). Η μέθοδος αυτή, παρέχει έναν έλεγχο για την υπόθεση ότι υπάρχει μια συνάρτηση $m(Cig)$ και τέσσερις σταθερές a_1, a_2, a_3, a_4 , τέτοια ώστε

$$E((Blad)^4 | Cig) = E((Blad)^4) + a_1 m(Cig)$$

$$E((Lung)^4 | Cig) = E((Lung)^4) + a_2 m(Cig)$$

$$E((Kid)^4 | Cig) = E((Kid)^4) + a_3 m(Cig)$$

$$E((Leuk)^4 | Cig) = E((Leuk)^4) + a_4 m(Cig).$$

Εφαρμόζουμε τη μέθοδο *SIR*, επιλέγοντας 14 slices, και παίρνουμε τα αποτελέσματα του Πίνακα 3.5. Παρατηρώντας τον πρώτο έλεγχο που δίνει η μέθοδος *SIR*, συμπεραίνουμε ότι το πρόβλημα έχει τουλάχιστον 1D δομή, αφού το *p-value* του ελέγχου είναι $0.008 < 0.05$. Από τον δεύτερο έλεγχο, που ελέγχει την υπόθεση ότι η παλινδρόμηση έχει το πολύ 1D δομή έναντι της εναλλακτικής υπόθεσης ότι έχει το λιγότερο 2D δομή, συμπεραίνουμε ότι επειδή το *p-value* ισούται με 0.131, που ξεπερνά την τιμή 0.05, η παλινδρόμηση έχει 1D δομή. Αυτό συμφωνεί με το συμπέρασμα που είχαμε βγάλει και με τη γραφική μέθοδο. Άρα η εξαρτημένη μεταβλητή *Cig* εξαρτάται από ένα μόνο γραμμικό συνδυασμό των μετασχηματισμένων μεταβλητών, δηλαδή των

$$h = 0.346(Blad)^4 + 0.936(Kid)^4 - 0.061(Leuk)^4 + 0.001(Lung)^4.$$

Στη συνέχεια θα μετασχηματίσουμε και την εξαρτημένη μεταβλητή *Cig*, υψώνοντάς τη εις την τετάρτη, και θα εφαρμόσουμε πάλι τη μέθοδο *SIR* για να δούμε αν αυτό θα αλλάξει τη δομική διάσταση του προβλήματος. Τα αποτελέσματα της μεθόδου εμφανίζονται στον Πίνακα 3.6. Παρατηρούμε ότι αντικαθιστώντας την εξαρτημένη μεταβλητή με το μετασχηματισμό της, τα αποτελέσματα δεν αλλάζουν σχεδόν καθόλου. Και πάλι συμπεραίνουμε ότι η δομική διάσταση του προβλήματος της παλινδρόμησης έχει 1D δομή και ο γραμμικός συνδυασμός είναι ο ίδιος με την προηγούμενη περίπτωση. Αυτό έρχεται να επιβεβαιώσει τον ισχυρισμό των Cook and Weisber (1994) ότι μονότονοι μετασχηματισμοί της εξαρτημένης μεταβλητής δεν επηρεάζουν τη δομική διάσταση του προβλήματος.

Inverse Regression SIR, Name of Dataset = smokingandcancer

Response = CIG

Predictors = (BLAD^4 KID^4 LEUK^4 LUNG^4)

Number of slices = 14

Slices sizes are: (4 4 3 3 3 4 3 3 3 3 3 3 3 2)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2		Lin Comb 3	
Predictors	Raw	Std.	Raw	Std.	Raw	Std.
BLAD^4	0.346	0.749	0.098	0.513	-0.553	-0.316
KID^4	0.936	0.324	-0.995	-0.836	0.408	0.037
LEUK^4	-0.061	-0.299	-0.011	-0.136	-0.726	-0.948
LUNG^4	0.001	0.495	-0.000	-0.138	-0.000	-0.002

Eigenvalues	0.779	0.451	0.389
R^2(OLS SIR)	0.971	0.971	0.999

Approximate Chi-squared test statistics based on partial sums of eigenvalues times 44

Number of Components	Test Statistic	df	p-value
1	79.893	52	0.008
2	45.597	36	0.131
3	25.769	22	0.262
4	8.6636	10	0.564

Πίνακας 3.5: Αποτελέσματα μεθόδου SIR με 14 slices

Inverse Regression SIR, Name of Dataset = smokingandcancer

Response = CIG^4

Predictors = (BLAD^4 KID^4 LEUK^4 LUNG^4)

Number of slices = 14

Slices sizes are: (4 4 3 3 3 4 3 3 3 3 3 3 2)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2		Lin Comb 3	
Predictors	Raw	Std.	Raw	Std.	Raw	Std.
BLAD^4	0.346	0.749	0.098	0.513	-0.553	-0.316
KID^4	0.936	0.324	-0.995	-0.836	0.408	0.037
LEUK^4	-0.061	-0.299	-0.011	-0.136	-0.726	-0.948
LUNG^4	0.001	0.495	-0.000	-0.138	-0.000	-0.002

Eigenvalues	0.779	0.451	0.389
R^2(OLS SIR)	0.898	0.934	0.965

Approximate Chi-squared test statistics based on partial sums of eigenvalues times 44

Number of Components	Test Statistic	df	p-value
1	79.893	52	0.008
2	45.597	36	0.131
3	25.769	22	0.262
4	8.6636	10	0.564

Πίνακας 3.6: Αποτελέσματα μεθόδου SIR με 14 slices και μετασχηματισμένη εξαρτημένη μεταβλητή

3.3 Ανάλυση προβλήματος Ais

3.3.1 Περιγραφή του προβλήματος

Το πρόβλημα, που θα αναλύσουμε στην ενότητα αυτή, αφορά δεδομένα που συνελέχθησαν από 102 αθλητές και 100 αθλήτριες από το Αυστραλιανό Ινστιτούτο Αθλητισμού. Συγκεκριμένα, στο αρχείο «ais.lsp» που θα χρησιμοποιήσουμε περιέχονται 14 μεταβλητές, των οποίων η

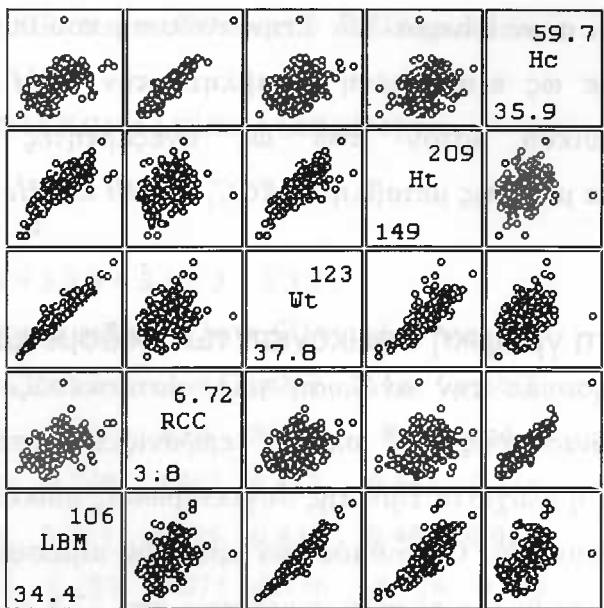
περιγραφή δίνεται στον Πίνακα 3.7. Στην ανάλυση που θα ακολουθήσει, θα χρησιμοποιήσουμε ως εξαρτημένη μεταβλητή την *LBM* που δηλώνει τη συγκέντρωση μυϊκού ιστού ενώ ως ανεξάρτητες μεταβλητές θα χρησιμοποιήσουμε μόνο τις μεταβλητές *RCC*, *Wt*, *Ht* και *Hc*.

3.3.2 Διδιάστατη γραφική απεικόνιση των δεδομένων

Θα ξεκινήσουμε την ανάλυσή μας, κατασκευάζοντας τον πίνακα διαγραμμάτων διασποράς, ο οποίος εμφανίζεται στο Σχήμα 3.14. Παρατηρούμε ότι η ελάχιστη τιμή της συγκέντρωσης μυϊκού ιστού είναι 34.4 ενώ η μέγιστη είναι 106. Ο αριθμός των ερυθρών αιμοσφαιρίων κυμαίνεται από 3.8 έως 6.72, το βάρος σε κιλά κυμαίνεται από 37.8 έως 123, το ύψος σε εκατοστά κυμαίνεται από 149 έως 209 ενώ ο αιματοκρίτης κυμαίνεται μεταξύ του 35.9 και του 59.7.

Όνομα μεταβλητής	Τύπος	Μέγεθος	Περιγραφή
% <i>Bfat</i>	Αριθμητική	202	Ποσοστό λίπους
<i>BMI</i>	Αριθμητική	202	Δείκτης μάζας σώματος
<i>Ferr</i>	Αριθμητική	202	Συγκέντρωση φερριτίνης του πλάσματος
<i>Hc</i>	Αριθμητική	202	Αιματοκρίτης
<i>Hg</i>	Αριθμητική	202	Αιμογλουμπίνη
<i>Ht</i>	Αριθμητική	202	Ύψος (εκατοστά)
<i>LBM</i>	Αριθμητική	202	Συγκέντρωση μυϊκού ιστού
<i>RCC</i>	Αριθμητική	202	Αριθμός ερυθρών αιμοσφαιρίων
<i>Sex</i>	Αριθμητική	202	Φύλο αθλητή
<i>SSF</i>	Αριθμητική	202	Μέτρηση υποδόριου λιπώδους ιστού
<i>WCC</i>	Αριθμητική	202	Αριθμός λευκών αιμοσφαιρίων
<i>Wt</i>	Αριθμητική	202	Βάρος (κιλά)
<i>Label</i>	Κείμενο	202	Ετικέτα
<i>Sport</i>	Κείμενο	202	Άθλημα

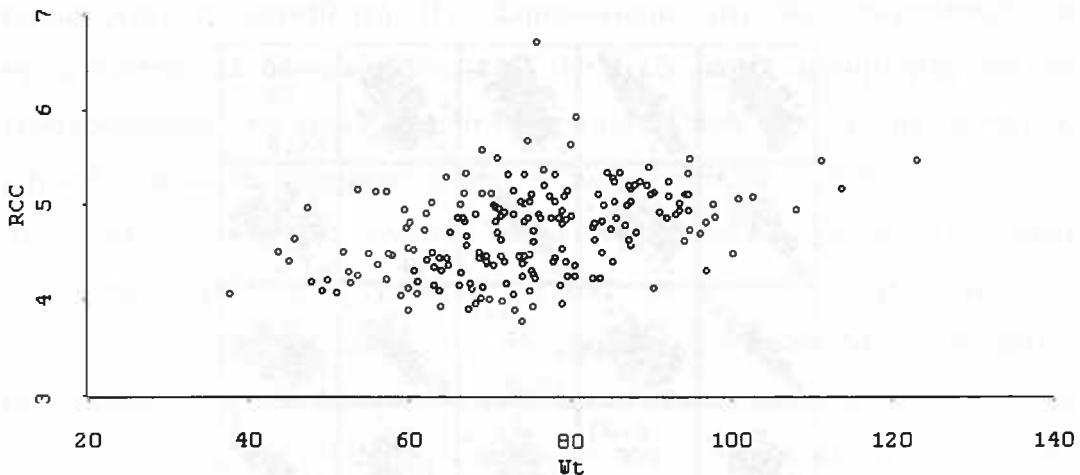
Πίνακας 3.7: Περιγραφή μεταβλητών προβλήματος *Ais*



Σχήμα 3.14: Πίνακας διαγραμμάτων διασποράς των δεδομένων *Ais*

Σημειώνουμε ότι τα διαγράμματα που βρίσκονται πάνω από την κύρια διαγώνιο του πίνακα, αποτελούν τις αντίστροφες απεικονίσεις (mirror images) των διαγραμμάτων που βρίσκονται κάτω από την κύρια διαγώνιο. Για παράδειγμα, τα γραφήματα που βρίσκονται στην πρώτη στήλη του πίνακα είναι οι αντίστροφες απεικονίσεις των γραφημάτων της τελευταίας γραμμής του πίνακα. Για παράδειγμα, το κάτω δεξιά διάγραμμα {*Hc*, *LBM*} είναι η αντίστροφη απεικόνιση του πάνω αριστερά διαγράμματος {*LBM*, *Hc*}.

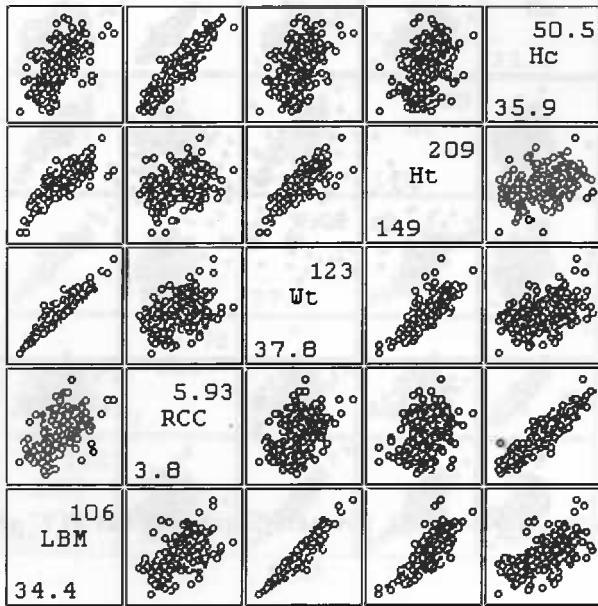
Παρατηρώντας τον πίνακα των διαγραμμάτων διασποράς, συμπεραίνουμε ότι όλα τα διαγράμματα διασποράς, αν και επηρεάζονται από κάποια απομακρυσμένα σημεία, εμφανίζονται να είναι γραμμικά και πιο συγκεκριμένα όλες οι μεταβλητές είναι θετικά συσχετισμένες. Η γραμμική σχέση θα μπορούσε να φανεί καλύτερα, αν μειώναμε το aspect ratio κάθε διαγράμματος, το οποίο όμως δεν εμφανίζεται εδώ για λόγους χώρου. Παρόλα αυτά δίνουμε στο Σχήμα 3.15 το διάγραμμα {*Wt*, *RCC*} με μικρότερο aspect ratio, από όπου μπορούμε να συμπεράνουμε ότι όντως υπάρχει μια γραμμική σχέση μεταξύ των δύο ανεξάρτητων μεταβλητών. Το ίδιο μπορεί να παρατηρηθεί και για τα υπόλοιπα διαγράμματα διασποράς του πίνακα που δεν δείχνουν ισχυρή γραμμικότητα.



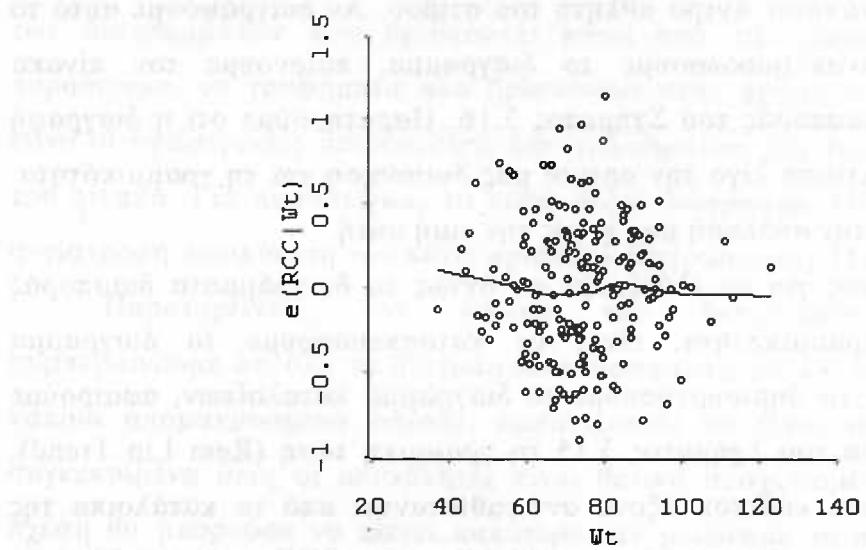
Σχήμα 3.15: Διάγραμμα διασποράς των μεταβλητών Wt και RCC με μικρότερο aspect ratio

Αναφέραμε νωρίτερα κατά το σχολιασμό των διαγραμμάτων ότι υπάρχει κάποιο απομονωμένο σημείο. Αυτό φαίνεται ευκρινώς στην πρώτη γραμμή του πίνακα και συνεπώς και στην τελευταία στήλη του. Το σημείο αυτό ανήκει σε κάποιον άντρα αθλητή του στίβου. Αν διαγράψουμε αυτό το σημείο και επανακλιμακώσουμε το διάγραμμα, παίρνουμε τον πίνακα διαγραμμάτων διασποράς του Σχήματος 3.16. Παρατηρούμε ότι η διαγραφή του σημείου, βελτίωσε λίγο την οπτική μας διαισθηση για τη γραμμικότητα. Ήα συνεχίσουμε την ανάλυσή μας χωρίς την τιμή αυτή.

Ένας τρόπος για να ελέγξουμε αν όντως τα διαγράμματα διασποράς παρουσιάζουν γραμμικότητα, είναι να κατασκευάσουμε το διάγραμμα καταλοίπων. Για να δημιουργήσουμε το διάγραμμα καταλοίπων, αφαιρούμε από το διάγραμμα του Σχήματος 3.15 τη γραμμική τάση (Rem Lin Trend), οπότε οι τιμές του κάθετου άξονα αντικαθίστανται από τα κατάλοιπα της παλινδρόμησης με εξαρτημένη μεταβλητή την RCC και ανεξάρτητη μεταβλητή την Wt . Στη συνέχεια προσαρμόζουμε στο διάγραμμα καταλοίπων μια ομαλή καμπύλη lowess. Από το Σχήμα 3.17, όπου εμφανίζεται το διάγραμμα καταλοίπων με προσαρμοσμένη μια lowess με παράμετρο $f = 0.6$ καμπύλη, συμπεραίνουμε ότι όντως το διάγραμμα παρουσιάζει γραμμικότητα. Αυτό συμβαίνει γιατί η καμπύλη lowess προσεγγίζει την ευθεία γραμμή που περνά από τη τιμή 0. Στο ίδιο συμπέρασμα καταλήγουμε αν ακολουθήσουμε την ίδια διαδικασία και στα υπόλοιπα διαγράμματα διασποράς.



Σχήμα 3.16: Πίνακας διαγραμμάτων διασποράς μετά τη διαγραφή του απομονωμένου σημείου

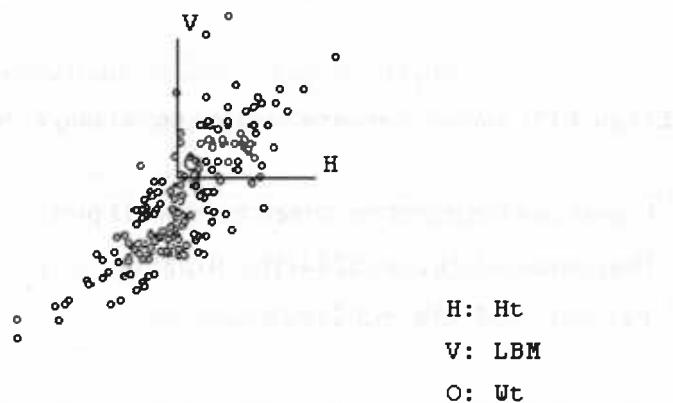


Σχήμα 3.17: Διάγραμμα καταλοίπων με προσαρμοσμένη *lowess* καμπύλη

3.3.3 Τρισδιάστατη γραφική απεικόνιση

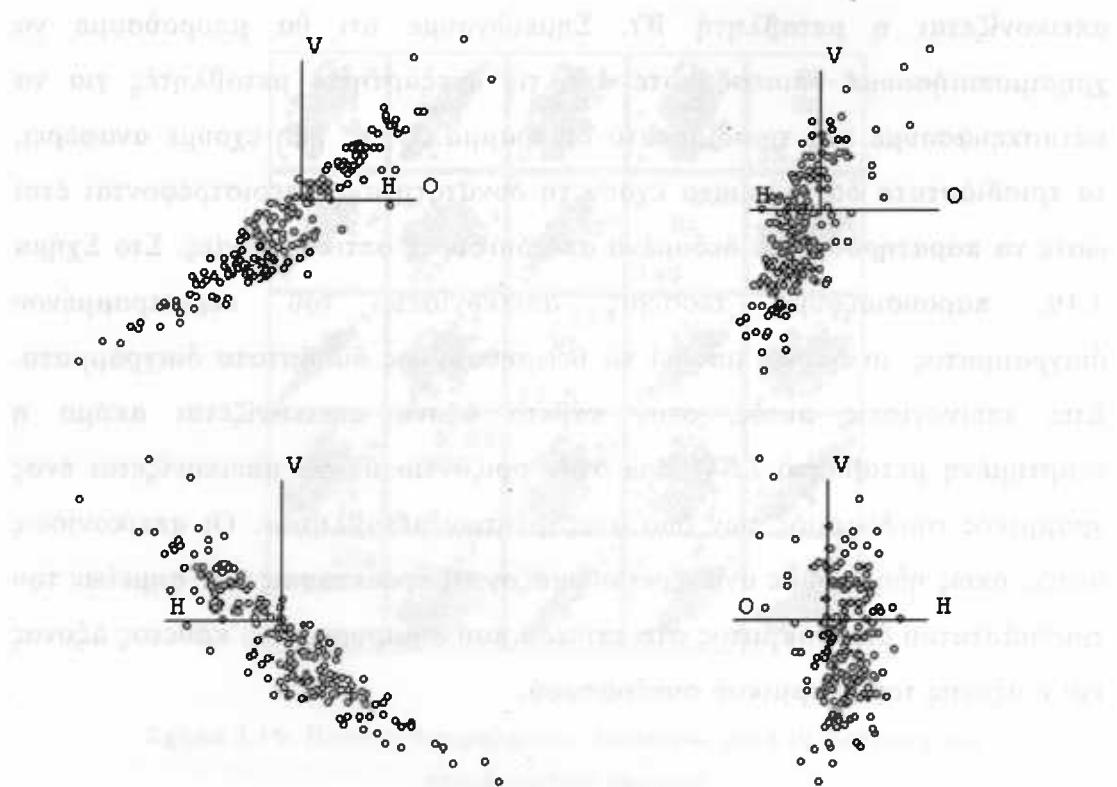
Στο Σχήμα 3.18 παρουσιάζουμε ένα τρισδιάστατο διάγραμμα, όπου στον κάθετο άξονα (V) απεικονίζεται η εξαρτημένη μεταβλητή LBM , στον οριζόντιο άξονα (H) η μεταβλητή Ht ενώ στον εκτός της οθόνης άξονα (O)

απεικονίζεται η μεταβλητή Wt . Σημειώνουμε ότι θα μπορούσαμε να χρησιμοποιήσουμε οποιεσδήποτε από τις ανεξάρτητες μεταβλητές για να κατασκευάσουμε ένα τρισδιάστατο διάγραμμα. Όπως ήδη έχουμε αναφέρει, τα τρισδιάστατα διαγράμματα έχουν τη δυνατότητα να περιστρέφονται έτσι ώστε να παρατηρούμε τα δεδομένα από διάφορες οπτικές γωνίες. Στο Σχήμα 3.19, παρουσιάζουμε τέσσερις απεικονίσεις του περιστραμμένου διαγράμματος, οι οποίες μπορεί να θεωρηθούν ως διδιάστατα διαγράμματα. Στις απεικονίσεις αυτές, στον κάθετο άξονα απεικονίζεται ακόμα η εξαρτημένη μεταβλητή LBM ενώ στον οριζόντιο άξονα απεικονίζεται ένας γραμμικός συνδυασμός των δυο ανεξάρτητων μεταβλητών. Οι απεικονίσεις αυτές, όπως ήδη έχουμε αναφέρει ονομάζονται προεκτάσεις των σημείων του τρισδιάστατου διαγράμματος στο επίπεδο που δημιουργούν ο κάθετος άξονας και ο άξονας του γραμμικού συνδυασμού.



Σχήμα 3.18: Τρισδιάστατο διάγραμμα μεταξύ των μεταβλητών Ht , LBM και Wt

Για παράδειγμα, ο γραμμικός συνδυασμός των μεταβλητών Ht και Wt που εμφανίζονται στην τελευταία διδιάστατη απεικόνιση του Σχήματος 3.19, είναι ο $h \approx 0.2441H - 0.158O$, όπου H είναι η μεταβλητή Ht και O είναι η μεταβλητή Wt . Τα αποτελέσματα που δίνει το πρόγραμμα *Arc* παρουσιάζονται στον Πίνακα 3.8. Από τον πίνακα αυτό παίρνουμε και την τιμή του d που είχαμε στη σχέση (1.3) ισούται περίπου με -3.101.



Σχήμα 3.19: Διάφορες απεικονίσεις του περιστραμμένου διαγράμματος

Linear combinations on screen axes in 3D plot.

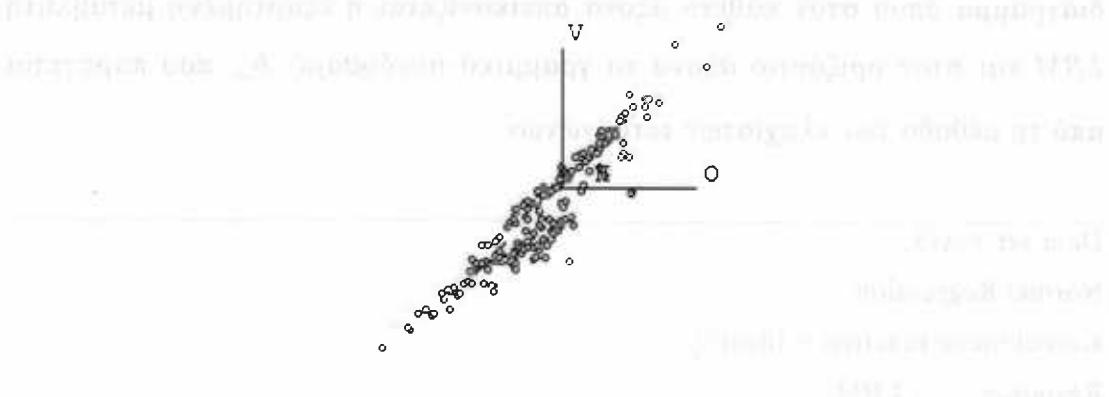
Horizontal: $-3.101 + .02441H - .01580O$

Vertical: $-1.959 + .02792V$

Πίνακας 3.8: Οι γραμμικοί συνδυασμού που αποτελούν τους άξονες του τρισδιάστατου διαγράμματος

3.3.4 Απεικόνιση της γραμμικής παλινδρόμησης

Στη συνέχεια, θα δούμε πως συμπεριφέρεται η συνάρτηση παλινδρόμησης $E(LBM | Ht, Wt)$. Ξεκινώντας από το τρισδιάστατο διάγραμμα του Σχήματος 3.18, το περιστρέφουμε έτσι ώστε αυτό να δείχνει όσο το δυνατόν την πιο ισχυρή γραμμική τάση (Σχήμα 3.20).



Σχήμα 3.20: Η διδιάστατη απεικόνιση με την πιο ισχυρή γραμμική τάση

Για τη διδιάστατη αυτή απεικόνιση, ο γραμμικός συνδυασμός των μεταβλητών Ht και Wt είναι ο $h^* \approx 0.00615H + 0.02301O$, όπου H είναι η μεταβλητή Ht και O είναι η μεταβλητή Wt (Πίνακας 3.9). Η διδιάστατη αυτή απεικόνιση λέμε ότι είναι το ιδανικό διάγραμμα περίληψης.

Linear combinations on screen axes in 3D plot.

Horizontal: $-2.954 + .00615H + .02301O$

Vertical: $-1.959 + .02792V$

Πίνακας 3.9: Οι γραμμικοί συνδυασμού που αποτελούν τους
άξονες του διαγράμματος του Σχήματος 3.20

Από τον Πίνακα 3.10, όπου παρουσιάζει τα αποτελέσματα της παλινδρόμησης της εξαρτημένης μεταβλητής LBM με το γραμμικό συνδυασμό h^* , παρατηρούμε ότι η σταθερά και η κλίση μέσω της μεθόδου ελαχίστων τετραγώνων είναι περίπου -29.2782 και 33.2283 , αντίστοιχα. Ισχύει δηλαδή,

$$E(LBM | Ht, Wt) = -29.2782 + 33.2283h^*.$$

Αυτό σημαίνει ότι $c^{-1} = 33.2283$ οπότε $c = 0.03009$. Τελικά έχουμε ότι

$$\begin{aligned} E(LBM | Ht, Wt) &= -29.2782 + 33.2283(0.00615Ht + 0.02301Wt) \\ &= -29.2782 + 0.204Ht + 0.7646Wt. \end{aligned}$$

Στη συνέχεια, θα εκτιμήσουμε τους άγνωστους συντελεστές παλινδρόμησης, μέσω της μεθόδου των ελαχίστων τετραγώνων. Ξεκινώντας από το αρχικό τρισδιάστατο διάγραμμα, κατασκευάζουμε το διδιάστατο

διάγραμμα όπου στον κάθετο άξονα απεικονίζεται η εξαρτημένη μεταβλητή LBM και στον οριζόντιο άξονα το γραμμικό συνδυασμό h_{ols} που παράγεται από τη μέθοδο των ελαχίστων τετραγώνων.

Data set = AIS,

Normal Regression

Kernel mean function = Identity

Response = LBM

Terms = (h^*)

Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
Constant	-29.2782	2.47421	-11.833	0.0000
h^*	33.2283	0.865926	38.373	0.0000

R Squared: 0.880418

Sigma hat: 4.53103

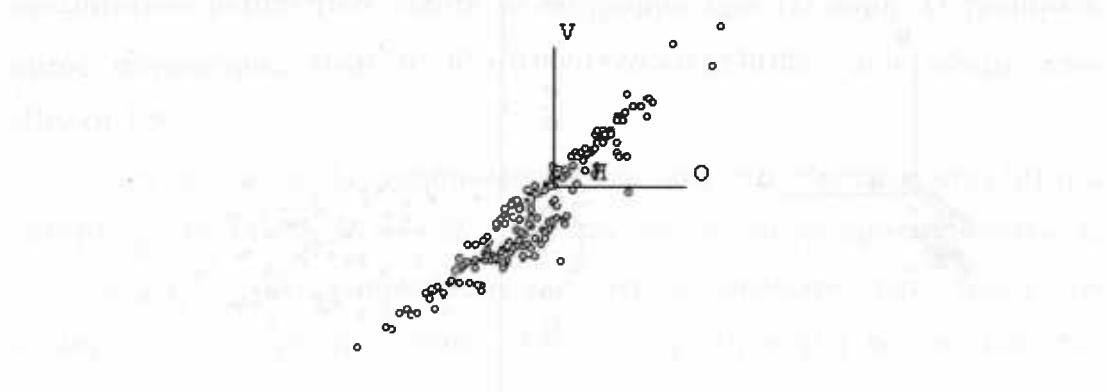
Number of cases: 202

Degrees of freedom: 200

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	1	30230.8	30230.8	1472.50	0.0000
Residual	200	4106.05	20.5303		
Lack of fit	199	4104.05	20.6234	10.31	0.2442
Pure Error	1	2.	2.		

Πίνακας 3.10: Αποτελέσματα παλινδρόμησης της LBM με την h^*

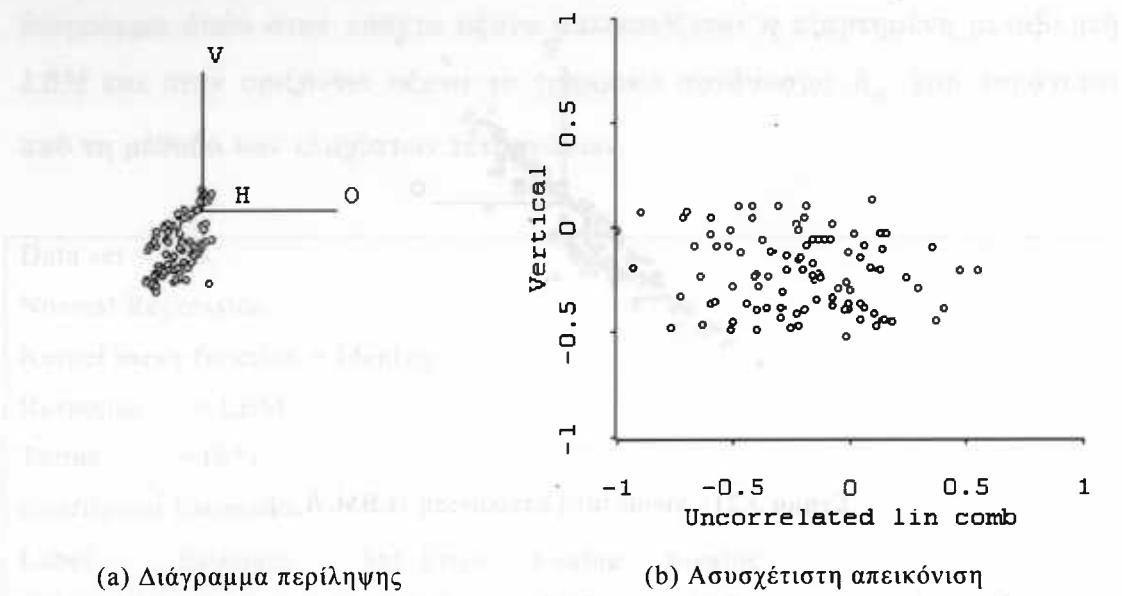


Σχήμα 3.21: Διδιάστατη απεικόνιση $\{LBM, h_{ols}\}$

Συγκρίνοντας τα Σχήματα 3.21 και 3.20, παρατηρούμε ότι είναι περίπου τα ίδια. Αυτό σημαίνει ότι η γραμμική τάση που επιλέξαμε με το μάτι στο Σχήμα 3.20, ήταν όντως από τις ισχυρότερες αν όχι η πιο ισχυρή.

Για να ελέγξουμε κατά πόσο το διάγραμμα περίληψης του Σχήματος 3.20 είναι ικανοποιητικό, ακολουθούμε τη διαδικασία της τμηματοποίησης. Ξεκινώντας από το Σχήμα 3.20, κατασκευάζουμε την ασυσχέτιστη απεικόνιση του τρισδιάστατου διαγράμματος και δημιουργούμε ένα slicer με ποσοστό 0.2. Στο Σχήμα 3.22 απεικονίζουμε ένα slice στο διάγραμμα περίληψης και τις αντίστοιχες παρατηρήσεις στην ασυσχέτιστη απεικόνιση του διαγράμματος περίληψης. Από τη συγκεκριμένη ασυσχέτιστη απεικόνιση, επειδή τα σημεία εμφανίζονται να βρίσκονται σε μια οριζόντια ζώνη, θεωρούμε ότι το διάγραμμα περίληψης είναι επαρκές.





Σχήμα 3.22: Διάγραμμα περίληψης και ασυσχέτιστη απεικόνιση αυτού

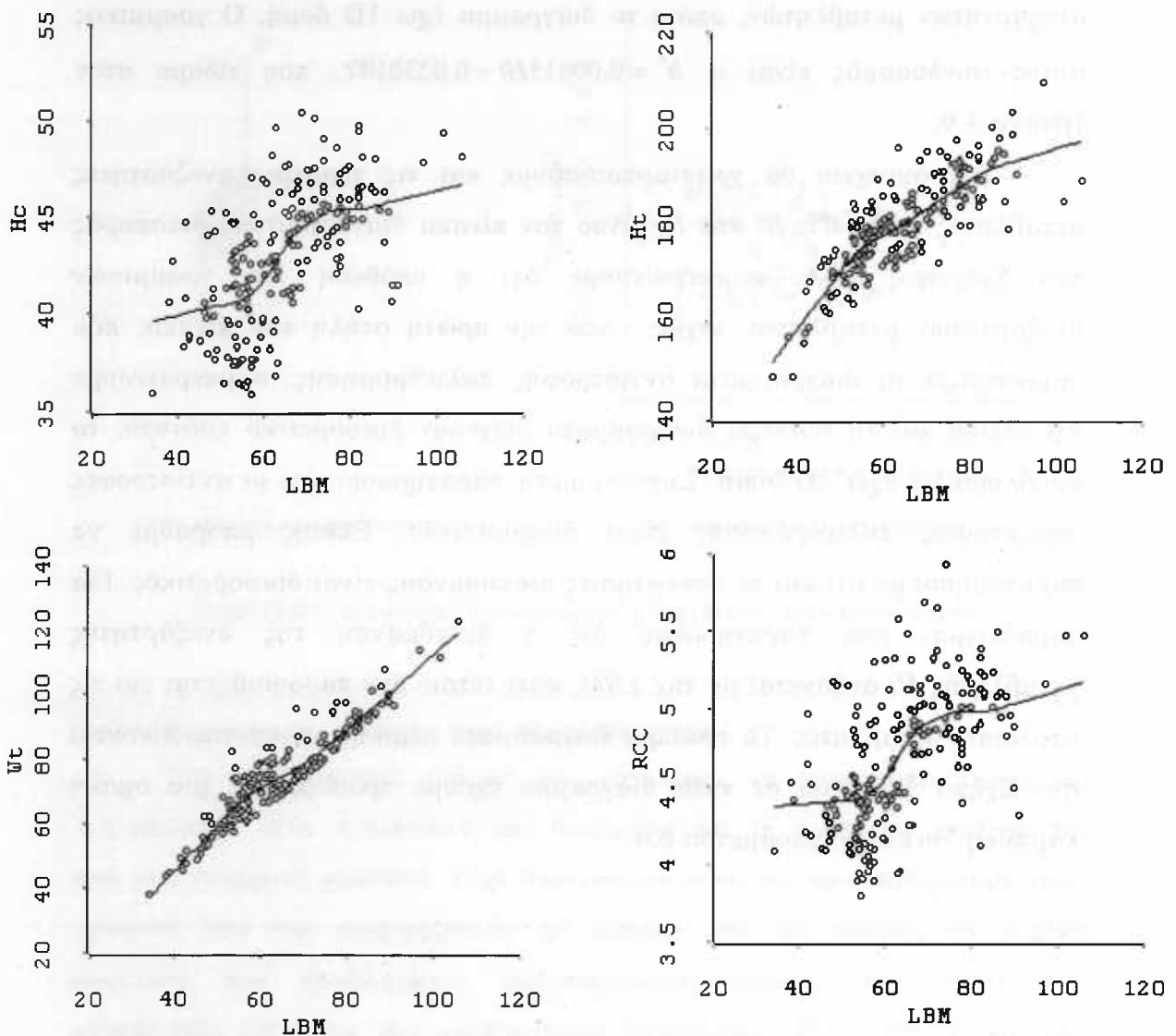
3.3.5. Εύρεση της δομής του προβλήματος

Στα προηγούμενα, είδαμε πώς μπορούμε να χρησιμοποιήσουμε τα διαγράμματα στην περίπτωση που θεωρούμε ότι τα δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Στην παράγραφο αυτή, θα προσπαθήσουμε τόσο γραφικά όσο και εφαρμόζοντας τη μέθοδο *SIR*, να βρούμε τη δομική διάσταση του προβλήματος παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής *LBM* και των ανεξάρτητων μεταβλητών *RCC*, *Wt*, *Ht* και *Hc*, χωρίς να υποθέτουμε την ισχύ κάποιου συγκεκριμένου μοντέλου. Να θυμίσουμε ότι το πρόβλημα αυτό, επειδή έχει τέσσερις ανεξάρτητες μεταβλητές μπορεί να έχει μέχρι 4D δομή.

Περιστρέφοντας και παρατηρώντας το τρισδιάστατο διάγραμμα του Σχήματος 3.18, δηλαδή το διάγραμμα $\{H_t, LBM, W_t\}$ συμπεραίνουμε ότι αυτό παρουσιάζει κάποιο συστηματικό πρότυπο. Συγκεκριμένα φαίνεται ότι η διακύμανση των παρατηρήσεων μεταβάλλεται και μάλιστα αυξάνεται. Αυτό οδηγεί στο συμπέρασμα ότι η υπόθεση της μηδενικής δομικής διάστασης δεν μπορεί να ισχύει. Συνεπώς το διάγραμμα θα έχει είτε 1D είτε 2D δομή. Για να βρούμε τι είδους δομή έχει, θα χρησιμοποιήσουμε τη μέθοδο με την ασυσχέτιστη απεικόνιση και το slicer. Επειδή όπως είδαμε και πριν, τα σημεία εμφανίζονται να βρίσκονται σε μια οριζόντια ζώνη, θεωρούμε ότι η μεταβλητή *LBM* εξαρτάται από ένα μόνο γραμμικό συνδυασμό των

ανεξάρτητων μεταβλητών, οπότε το διάγραμμα έχει 1D δομή. Ο γραμμικός αυτός συνδυασμός είναι ο $h^* \approx 0.00615Ht + 0.02301Wt$, που είδαμε στον Πίνακα 3.9.

Στη συνέχεια θα χρησιμοποιήσουμε και τις τέσσερις ανεξάρτητες μεταβλητές RCC , Wt , Ht και Hc . Από τον πίνακα διαγραμμάτων διασποράς του Σχήματος 3.16, συμπεραίνουμε ότι η υπόθεση των γραμμικών ανεξάρτητων μεταβλητών ισχύει. Από την πρώτη στήλη του πίνακα, που παρουσιάζει τα διαγράμματα αντίστροφης παλινδρόμησης, συμπεραίνουμε ότι επειδή και τα τέσσερα διαγράμματα δείχνουν διαφορετικό πρότυπο, το πρόβλημα θα έχει 2D δομή. Συγκεκριμένα παρατηρούμε ότι οι αντίστροφες συναρτήσεις παλινδρόμησης είναι διαφορετικές. Επίσης μπορούμε να παρατηρήσουμε ότι και οι συναρτήσεις διακύμανσης είναι διαφορετικές. Για παράδειγμα, ενώ παρατηρούμε ότι η διακύμανση της ανεξάρτητης μεταβλητής Ht αυξάνεται με την LBM , κάτι τέτοιο δεν παρουσιάζεται για τις υπόλοιπες μεταβλητές. Τα τέσσερα διαγράμματα μερικής απόκρισης, δίνονται στο Σχήμα 3.23 ενώ σε κάθε διάγραμμα έχουμε προσαρμόσει μια ομαλή καμπύλη *lowess* με παράμετρο 0.6.



Σχήμα 3.23: Διαγράμματα αντίστροφης παλινδρόμησης με προσαρμοσμένη *lowess*

Επειδή η εύρεση της δομικής διάστασης της παλινδρόμησης μέσω διαγραμμάτων βασίζεται πολύ σε οπτικές απεικονίσεις, θα εφαρμόσουμε και τη μέθοδο της τμηματικής αντίστροφης παλινδρόμησης (*SIR*). Η μέθοδος αυτή, παρέχει έναν έλεγχο για την υπόθεση ότι υπάρχει μια συνάρτηση $m(LBM)$ και τέσσερις σταθερές a_1, a_2, a_3, a_4 , τέτοια ώστε

$$E(RCC | LBM) = E(RCC) + a_1 m(LBM)$$

$$E(Wt | LBM) = E(Wt) + a_2 m(LBM)$$

$$E(Ht | LBM) = E(Ht) + a_3 m(LBM)$$

$$E(Hc | LBM) = E(Hc) + a_4 m(LBM).$$

Εφαρμόζουμε τη μέθοδο *SIR*, επιλέγοντας 25 slices, και παίρνουμε τα αποτελέσματα του Πίνακα 3.11. Παρατηρώντας τον πρώτο έλεγχο που δίνει η μέθοδος *SIR*, συμπεραίνουμε ότι το πρόβλημα έχει τουλάχιστον 1D δομή, αφού το *p-value* του ελέγχου είναι $0 < 0.05$. Από τον δεύτερο έλεγχο, που ελέγχει την υπόθεση ότι η παλινδρόμηση έχει το πολύ 1D δομή έναντι της εναλλακτικής υπόθεσης ότι έχει το λιγότερο 2D δομή, συμπεραίνουμε ότι επειδή το *p-value* ισούται με 0.003, που και πάλι είναι μικρότερο από την τιμή 0.05, η παλινδρόμηση έχει το λιγότερο 1D δομή. Από τον τρίτο έλεγχο για την υπόθεση ότι η παλινδρόμηση έχει το πολύ 2D δομή έναντι της εναλλακτικής υπόθεσης ότι έχει το λιγότερο 3D δομή, παρατηρούμε ότι το *p-value* ξεπερνά την τιμή 0.05, οπότε το πρόβλημα έχει 2D δομή. Άρα η εξαρτημένη μεταβλητή *LBM* εξαρτάται από δυο γραμμικούς συνδυασμούς των μετασχηματισμένων μεταβλητών, δηλαδή τους

$$h_1 = -0.762RCC + 0.371Wt + 0.175Ht + 0.501Hc$$

και

$$h_2 = 0.995RCC - 0.053Wt + 0.034Ht + 0.083Hc.$$

Ολοκληρώνοντας την ανάλυση των δεδομένων, να σημειώσουμε ότι όταν χρησιμοποιήσαμε τις τέσσερις ανεξάρτητες μεταβλητές *RCC*, *Wt*, *Ht* και *Hc*, η δομική διάσταση του προβλήματος ήταν 2D ενώ στην περίπτωση που ελέγχαμε τη δομή της παλινδρόμησης μέσω του τρισδιάστατου διαγράμματος $\{Ht, LBM, Wt\}$, βρήκαμε 1D δομή. Στον Πίνακα 3.12, παρουσιάζονται τα αποτελέσματα της μεθόδου *SIR* χρησιμοποιώντας μόνο τις μεταβλητές *Ht* και *Wt*. Συμπεραίνουμε ότι επειδή το *p-value* του πρώτου ελέγχου είναι μικρότερο του 0.05 ενώ του δεύτερου ελέγχου είναι μεγαλύτερο του 0.05, τότε το πρόβλημα έχει μοναδιαία δομική διάσταση, πράγμα που επιβεβαιώνει την απόφασή μας μέσω του τρισδιάστατου διαγράμματος. Άρα η ανεξάρτητη μεταβλητή *LBM* εξαρτάται από έναν γραμμικό συνδυασμό των *Ht* και *Wt* που είναι ο

$$h = 0.925Wt + 0.380Ht.$$

Αυτό σημαίνει ότι η εισαγωγή νέων μεταβλητών σε ένα πρόβλημα παλινδρόμησης μπορεί να αλλάξει τη δομική διάσταση του προβλήματος.

Inverse Regression SIR; Name of Dataset = AIS

Response = LBM

Predictors = (RCC Wt Ht Hc)

Number of slices = 23

Slices sizes are: (9 9 8 9 8 8 8 8 8 8 9 9 12 8 10 12 9 9 8 10 8 7)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2		Lin Comb 3	
Predictors	Raw	Std.	Raw	Std.	Raw	Std.
RCC	-0.762	-0.061	0.995	0.469	0.994	0.642
Wt	0.371	0.898	-0.053	-0.754	0.017	0.339
Ht	0.175	0.296	0.034	0.339	-0.031	-0.426
Hc	0.501	0.319	0.083	0.312	-0.104	-0.539

Eigenvalues	0.918	0.281	0.112
R^2(OLS SIR)	0.998	0.998	1.000

Approximate Chi-squared test statistics based on partial
sums of eigenvalues times 202

Number of Components	Test Statistic	df	p-value
1	283.02	88	0.000
2	97.521	63	0.003
3	40.736	40	0.438
4	18.198	19	0.509

Πίνακας 3.11: Αποτελέσματα μεθόδου SIR με 25 slices

Inverse Regression SIR, Name of Dataset = AIS

Response = LBM,

Predictors = (Wt Ht)

Number of slices = 23

Slices sizes are: (9 9 8 9 8 8 8 8 8 9 9 12 8 10 12 9 9 8 10 8 7)

Std. coef. use predictors scaled to have SD equal to one.

	Lin Comb 1		Lin Comb 2	
Predictors	Raw	Std.	Raw	Std.
Wt	0.925	0.961	-0.521	-0.657
Ht	0.380	0.276	0.854	0.754

Eigenvalues 0.895 0.128

R^2(OLS| SIR) 1.000 1.000

Approximate Chi-squared test statistics based on partial
sums of eigenvalues times 202

Number of Components	Test Statistic	df	p-value
1	206.67	44	0.000
2	25.952	21	0.208

Πίνακας 3.12: Αποτελέσματα μεθόδου SIR με 25 slices και ανεξάρτητες

μεταβλητές τις Wt και Ht

1. Generalized linear model of the number of cases per day	1.601 ± 0.002
2. Logistic regression	0.00167 ± 0.00002

Logistic regression: 1.6

Generalized linear model:

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

$$1.601 \pm 0.002 = 1.601 \pm 0.002$$

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Generalized linear model: 1.601 ± 0.002. The difference between the two models is statistically significant at the level of 0.001 (see Fig. 1).

Κεφάλαιο 4:

Συμπεράσματα

4.1 Συνοπτική παρουσίαση

Κύριο αντικείμενο της παρούσας διπλωματικής εργασίας ήταν να παρουσιάσει τους τρόπους με τους οποίους μπορούμε να ελέγξουμε τη δομική διάσταση σε προβλήματα παλινδρόμησης. Με άλλα λόγια, να κατανοήσουμε τον τρόπο με τον οποίο η εξαρτημένη μεταβλητή εξαρτάται από μια ή περισσότερες ανεξάρτητες (ελεγχόμενες) τυχαίες μεταβλητές.

Οι τρόποι τους οποίους χρησιμοποιούμε για να ανακαλύψουμε τη δομική διάσταση προβλημάτων παλινδρόμησης είναι η χρήση διαγραμμάτων και αριθμητικών μεθόδων.

4.1.1 Γραφική μέθοδος

Για να βρούμε γραφικά τη διάσταση ενός προβλήματος παλινδρόμησης, θα πρέπει να υποθέσουμε ότι ισχύει η υπόθεση της 1D δομής. Στη συνέχεια, για να ελέγξουμε την ανεξαρτησία των ερμηνευτικών μεταβλητών και να επιβεβαιώσουμε ότι πράγματι ισχύει η υπόθεση της 1D δομής, κατασκευάζουμε ένα $(p+1)$ -διάστατο γράφημα των p ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής. Κατόπιν, αντιστρέφουμε το πρόβλημα και αντί να μελετήσουμε την κατανομή της $y|x$, μελετάμε την αντίστροφη παλινδρόμηση, έτσι ώστε να δούμε τον τρόπο με τον οποίο συμπεριφέρεται η κατανομή του διανύσματος x δοθέντων των τιμών της μεταβλητής y , δηλαδή η κατανομή του $x|y$. Στην περίπτωση αυτή, έχουμε να αντιμετωπίσουμε p προβλήματα απλής παλινδρόμησης που μπορούν να

μελετηθούν με ένα απλό διδιάστατο διάγραμμα διασποράς το καθένα, της μορφής $\{y, x_j\}, j = 1, \dots, p$, τα οποία ονομάζονται αντίστροφα διαγράμματα μερικής απόκρισης. Αν πράγματι έχουμε 1D δομή τα p αντίστροφα διαγράμματα μερικής απόκρισης θα πρέπει να έχουν την ίδια μορφή. Αν κάτι τέτοιο δεν ισχύει, τότε θα πρέπει να εγκαταλείψουμε την υπόθεση της 1D δομής. Επίσης, στα διαγράμματα αυτά η διακύμανση θα πρέπει να αλλάζει με τον ίδιο τρόπο. Αν αυτό δεν συμβαίνει τότε δεν ισχύει η υπόθεση της 1D δομής. Η μόνη εξαίρεση είναι όταν έχουμε ανεξάρτητες μεταβλητές. Για να είναι συνεπή με την 1D δομή, τα διαγράμματα που δείχνουν μη εξάρτηση της συνάρτησης αντίστροφης παλινδρόμησης από το y , θα πρέπει να δείχνουν ανεξαρτησία και της συνάρτησης αντίστροφης διακύμανσης ακόμη και αν η διακύμανση δεν είναι σταθερή σε άλλα διαγράμματα.

Σε καθένα από τα p αντίστροφα διαγράμματα μερικής απόκρισης, μπορούμε να προσαρμόσουμε έναν εξομαλυντή, ο οποίος αποτελεί εκτίμηση της συνάρτησης αντίστροφης παλινδρόμησης $E(x_j | y)$. Εάν ισχύει η 1D δομή, η συνάρτηση αντίστροφης παλινδρόμησης θα πρέπει να προσεγγίζει την ποσότητα $E(x_j) + a_j m(y)$.

4.1.2 Αριθμητικές μέθοδοι

Όταν χρησιμοποιούμε τη γραφική μέθοδο για να βρούμε τις διαστάσεις ενός προβλήματος παλινδρόμησης, στηριζόμαστε κατά κύριο λόγο σε οπτικές εντυπώσεις. Για το λόγο αυτό, έχουν προταθεί τρεις αριθμητικές μέθοδοι: η τμηματική αντίστροφη παλινδρόμηση, η μέθοδος SAVE και η μέθοδος Principal Hessian directions.

Τη μέθοδο της τμηματικής αντίστροφης παλινδρόμησης πρότειναν οι Duan and Li (1991) και Li (1991). Η μέθοδος αυτή υποθέτει ότι οι ερμηνευτικές μεταβλητές ακολουθούν την κανονική κατανομή και παρέχει έναν έλεγχο για την ισχύ ή όχι της 1D δομής. Στη μέθοδο αυτή, η συνάρτηση παλινδρόμησης σε κάθε διάγραμμα αντίστροφης παλινδρόμησης εκτιμάται μέσω εξομάλυνσης. Ο Li (1991) έδειξε ότι οι εκτιμημένες ομαλές καμπύλες μπορούν να συγκριθούν για να δώσουν ένα τεστ για τη διάσταση του

προβλήματος. Αυτό που περιμένουμε είναι ότι οι ομαλές καμπύλες στα αντίστροφα διαγράμματα μερικής απόκρισης θα είναι ίδιες εκτός από έναν παράγοντα κλίμακας που εφαρμόζεται σε κάθε μια. Για να το ελέγξουμε αυτό, χρησιμοποιούμε μια ομαλή καμπύλη που βασίζεται στο χωρισμό των δεδομένων σε μη επικαλυπτόμενα slices σύμφωνα με τη μεταβλητή απόκρισης y . Στη συνέχεια υπολογίζουμε το μέσο όρο των x , σε κάθε slice ενώ ο αριθμός των slices είναι η σταθερά που ρυθμίζει την ομαλή καμπύλη.

Κατόπιν, ο αλγόριθμος της μεθόδου αυτής, κατασκευάζει ένα τρισδιάστατο γράφημα της μορφής $\{h_1, y, h_2\}$, όπου τα $h_1 = \mathbf{b}^T \mathbf{x}$ και $h_2 = \mathbf{a}^T \mathbf{x}$ είναι γραμμικοί συνδυασμοί των ανεξάρτητων μεταβλητών. Αν η 1D δομή είναι πραγματικά κατάλληλη, τότε η διδιάστατη απεικόνιση $\{h_1, y\}$ πριν από την περιστροφή θα αποτελεί το διάγραμμα περίληψης που εκτιμάται από τη μέθοδο SIR. Αν χρειάζεται δομή 2D τότε απαιτείται και δεύτερος γραμμικός συνδυασμός, ο οποίος είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών h_2 στον εκτός σελίδας άξονα. Αν ισχύει η 2D δομή, τότε το πλήρες τρισδιάστατο γράφημα $\{h_1, y, h_2\}$ θα αποτελεί το διάγραμμα περίληψης που προκύπτει από τη μέθοδο τμηματικής αντίστροφης παλινδρόμησης

Η μέθοδος SAVE, προτάθηκε από τους Cook and Weisberg (1991), και είναι παρόμοια με τη μέθοδο SIR. Η μέθοδος αυτή χρησιμοποιεί τόσο τη συνθήκη γραμμικότητας της συνάρτησης παλινδρόμησης όσο και τη συνθήκη της σταθερής συνάρτησης διακύμανσης για να βρει την αντίστροφη δομή.

Οι Cook and Weisberg (1991), οι οποίοι εφάρμοσαν τη μέθοδο της τμηματικής αντίστροφης παλινδρόμησης (SIR) σε ένα παράδειγμα, κατέληξαν στο συμπέρασμα ότι αυτή δεν καταλήγει πάντα σε σωστή απόφαση σχετικά με τις διαστάσεις σε ένα πρόβλημα παλινδρόμησης και αυτό συμβαίνει όταν η διακύμανση των σφαλμάτων ϵ είναι μικρή. Αντίθετα, όταν η διακύμανση των σφαλμάτων είναι μεγάλη η μέθοδος SIR βρίσκει πάντα τη σωστή λύση. Ένα δεύτερο μειονέκτημα της μεθόδου SIR, σύμφωνα με τους Cook and Weisberg (1991), είναι η ανικανότητά που αυτή έχει στο να διαγνώσει τη συμμετρική εξάρτηση, όταν η μέση τιμή των τυποποιημένων τιμών του x σε κάθε slice είναι κοντά στο μηδέν

Η μέθοδος Principal Hessian directions για την εύρεση αντίστροφων δομών προτάθηκε επίσης από τον Li (1992). Η μέθοδος αυτή επίσης παρέχει έναν έλεγχο για τη διάσταση του προβλήματος παλινδρόμησης. Στη συνέχεια, η μέθοδος επεκτάθηκε από τον Cook (1998) και, όπως η μέθοδος *SAVE*, απαιτεί τόσο τη συνθήκη της γραμμικότητας της συνάρτησης παλινδρόμησης όσο και τη συνθήκη της σταθερής συνάρτησης διακύμανσης.

Η μέθοδος αυτή ως μέθοδος μείωσης της διάστασης ενός προβλήματος παλινδρόμησης θεωρείται καλύτερη όταν οι ανεξάρτητες μεταβλητές αντιστοιχούν σε μη γραμμικές τάσεις. Για το λόγο αυτό η μέθοδος φαίνεται να είναι καλύτερη όταν χρησιμοποιείται ως διαγνωστικός έλεγχος για κάποιο μοντέλο, αντικαθιστώντας την εξαρτημένη μεταβλητή y με τα κατάλοιπα e από ένα προσαρμοσμένο γραμμικό μοντέλο. Όταν χρησιμοποιείται κατά αυτόν τον τρόπο, ο αριθμός των γραμμικών συνδυασμών που προκύπτουν είναι ο μικρότερος αριθμός των επαρκών ανεξάρτητων μεταβλητών από την παλινδρόμηση των καταλοίπων e με το διάνυσμα των ανεξάρτητων μεταβλητών x .

4.2 Κριτική θεώρηση των μεθόδων

Στην παράγραφο αυτή, θα προσπαθήσουμε να κάνουμε μια σύντομη κριτική των παραπάνω μεθόδων. Όπως ήδη έχουμε αναφέρει, η γραφική μέθοδος στηρίζεται κατά κύριο λόγο σε οπτικές εντυπώσεις, πράγμα που την καθιστά υποκειμενική. Όσον αφορά τις αριθμητικές μεθόδους, αν και όλες δίνουν χρήσιμα αποτελέσματα στην πράξη, θα ήταν παράλογο να μην παρουσιάζουν κάποια μειονεκτήματα.

Ξεκινώντας από τη μέθοδο *SIR*, μπορούμε να πούμε ότι είναι καλύτερη από τις υπόλοιπες στην περίπτωση ύπαρξης γραμμικών τάσεων στη συνάρτηση παλινδρόμησης $E(y|x)$. Επιπλέον, η μέθοδος αυτή, δίνει καλύτερα αποτελέσματα όταν η συνάρτηση διακύμανσης $\text{var}(y|x)$ δεν είναι σταθερή. Το βασικό όμως μειονέκτημα της μεθόδου *SIR*, είναι ότι παρουσιάζεται γενικά αναποτελεσματική στην αναγνώριση της καμπυλότητας στη συνάρτηση παλινδρόμησης των καταλοίπων $E(e|x)$.

Όσον αφορά τη μέθοδο *SAVE*, αυτή βρίσκει πάντοτε εκτός από τις επαρκείς μεταβλητές που βρίσκουν οι άλλες δυο μέθοδοι, και όποιες άλλες υπάρχουν στην παλινδρόμηση. Αυτό όμως έχει και κάποιο κόστος καθώς η μέθοδος *SAVE* ψάχνει σε μια μεγαλύτερη κλάση γραμμικών συνδυασμών που περιλαμβάνει τη μικρότερη κλάση στην οποία ψάχνουν οι μέθοδοι *SIR* και *rHd*. Επίσης, αρκετές φορές η μέθοδος *SAVE*, αντιμετωπίζει δυσκολίες στο να βρει σχετικά ευθείες δομές ενώ οι άλλες δυο μέθοδοι τις βρίσκουν εύκολα. Αυτό συνήθως συμβαίνει όταν ο αριθμός των ανεξάρτητων μεταβλητών είναι μεγάλος.

Η μέθοδος *rHd*, είναι μια μέθοδος που δίνει καλύτερα αποτελέσματα στην αναγνώριση της καμπυλότητας στη συνάρτηση παλινδρόμησης των καταλοίπων $E(e | \mathbf{x})$. Στην περίπτωση όμως της εύρεσης των κατευθύνσεων όταν η συνάρτηση διακύμανσης των καταλοίπων $\text{var}(e | \mathbf{x})$ δεν είναι σταθερή, μειονεκτεί έναντι των μεθόδων *SIR* και *SAVE*.

Με βάση την παραπάνω κριτική ανασκόπηση των μεθόδων, προκύπτει ότι η μέθοδος της τμηματικής αντίστροφης παλινδρόμησης, θεωρείται ως καλύτερη και πιο αποτελεσματική. Άλλωστε είναι και η πιο διαδεδομένη και περισσότερο χρησιμοποιούμενη στην πράξη.



Βιβλιογραφία

A. ΞΕΝΗ

- Cook, R.D. (2000).** SAVE: a method for dimension reduction and graphics in regression, *Communications in Statistics: Theory Methods*, 29, 2109-2121
- Cook, R.D. (1998).** *Regression Graphics: Ideas for studying regressions thru graphics*. Wiley, New York
- Cook, R.D. and Weisberg, S. (1994).** *An Introduction to Regression Graphics*. Wiley, New York
- Cook, R.D. and Weisberg, S. (1994).** Discussion of “Sliced inverse regression for dimension reduction”, *J. Amer. Statist. Assoc.*, 86, 328-333
- Cook, R.D. and Weisberg, S. (1999).** *Applied Regression Including Computing and Graphics*. Wiley, New York
- Duan, N. and Li, K.C. (1991).** Slicing regression: a link-free regression method, *The Annals of Statistics*, 19, 2, 505-530
- Li, K.C. (1991).** Sliced inverse regression for dimension reduction (with discussion), *J. Amer. Statist. Assoc.*, 86, 316-342
- Li, K.C. (1992).** On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *J. Amer. Statist. Assoc.*, 87, 1025–39.

Biblioabafia

A. EINH

- Cook, R.D. (2000). SVAE: a method for dimension reduction and linkages in classification. *Communication in Statistics: Theory and Methods*, 29, 2103-2131.
- Cook, R.D. (1988). Regression Graphics. *Techniques for Transforming Data to Insights*. Cambridge, MA: John Wiley.
- Cook, R.D. and Weisberg, S. (1994). An Introduction to Regression Graphics. New York: Wiley.
- Cook, R.D. and Weisberg, S. (1994). Dimension of "Sliced inverse regression for dimension reduction". *Journal of American Statistical Association*, 89, 358-363.
- Cook, R.D. and Weisberg, S. (1990). Thruway Miles: some comments. *Comments on Current Books*, 1, 1-10.
- Duan, N. and Li, K.C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, 19, 5, 205-230.
- Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Annals of Statistics*, 19, 3-28.
- Li, K.C. (1995). On bicubic least squares dimension for data visualisation and dimension reduction: supports application of Sliced Inverse Regression. *Annals of Statistics*, 23, 1052-1066.



Δωρίδης

