Κεφάλαιο 18

Μη Παραμετρική Παλινδρόμηση

Το παραδοσιακό παραμετρικό μοντέλο δίνεται από την εξίσωση

$$y_i = f(\boldsymbol{\beta}, \mathbf{x_i'}) + \varepsilon_i,$$

όπου $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ το διάνυσμα των παραμέτρων που θα εκτιμηθούν, και $\mathbf{x_i'} = (x_1, \ldots, x_k)$ το διάνυσμα των επεξηγηματικών μεταβλητών για την i παρατήρηση. Για τα σφάλματα ε_i υποθέτουμε ότι είναι ανεξάρτητα και ακολουθούν την κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση σ^2. Η συνάρτηση $f(.)$ ορίζει την σχέση μεταξύ της μέσης τιμής της εξαρτημένης μεταβλητής και των επεξηγηματικών μεταβλητών. Το γενικό μη παραμετρικό μοντέλο παλινδρόμησης γράφεται με παρόμοιο τρόπο χωρίς όμως να ορίζεται η f:

$$y_i = f(\mathbf{x_i'}) + \varepsilon_i = f(x_{i1}, \ldots, x_{ik}) + \varepsilon_i.$$

Επιπρόσθετα, ο σκοπός της μη παραμετρικής παλινδρόμησης είναι να εκτιμήσει την συνάρτηση παλινδρόμησης f απ' ευθείας παρά να εκτιμήσει παραμέτρους. Οι περισσότερες μέθοδοι μη παραμετρικής παλινδρόμησης υποθέτουν ότι η f είναι ομαλή και συνεχής.

Ειδική περίπτωση του γενικού μοντέλου είναι η μη παραμετρική απλή παλινδρόμηση, στην οποία υπάρχει μόνο μία επεξηγηματική μεταβλητή:

$$y_i = f(x_i) + \varepsilon_i$$

Το μοντέλο αυτό ονομάζεται επίσης και «εξομάλυνση διαγράμματος διασποράς»

αφού κατασκευάζει μια ομαλή καμπύλη για το διάγραμμα διασποράς του y συναρτήσει του x.

Λόγω της δυσκολία εφαρμογής και απεικόνισης ενός γενικού μη παραμετρικού μοντέλου παλινδρόμησης με πολλές επεξηγηματικές μεταβλητές, έχουν αναπτυχθεί πιο περιοριστικά μοντέλα, όπως π.χ. το αθροιστικό μοντέλο παλινδρόμησης (additive regression model)

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_k(x_{ik}) + \varepsilon_i.$$

Για αυτό το μοντέλο υποθέτουμε ότι οι μερικές συναρτήσεις παλινδρόμησης $f_j(.)$ είναι ομαλές και μπορούν να εκτιμηθούν από τα δεδομένα. Το μοντέλο αυτό είναι πιο περιοριστικό από το γενικό μη παραμετρικό μοντέλο, αλλά λιγότερο περιοριστικό από το μοντέλο γραμμικής παλινδρόμησης, στο οποίο όλες οι μερικές συναρτήσεις παλινδρόμησης θεωρούνται γραμμικές.

18.1 Τοπική Πολυωνυμική Παλινδρόμηση

Απλή Παλινδρόμηση

Το μοντέλο απλής παλινδρόμησης το οποίο θεωρείται δίνεται από

$$y_i = f(x_i) + \varepsilon_i,$$

όπου $f(.)$ η άγνωστη παράμετρος. Η συνάρτηση παλινδρόμησης f θα εκτιμηθεί σε ένα συγκεκριμένο σημείο x_0. Αυτό είναι δυνατόν χρησιμοποιώντας τη πολυωνυμική παλινδρόμηση p-τάξης της y πάνω στη x των τοπικά σταθμισμένων ελαχίστων τετραγώνων (weighted least squares),

$$y_i = \alpha + b_1(x_i - x_0) + b_2(x_i - x_0)^2 + \ldots + +b_p(x_i - x_0)^p + e_i$$

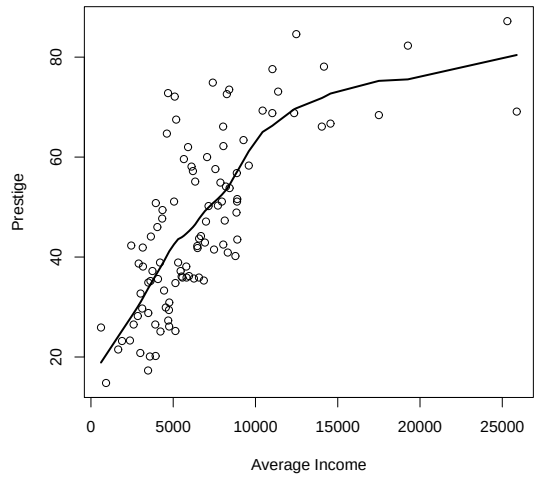
με την οποία οι παρατηρήσεις σταθμίζονται σε σχέση με το πόσο κοντά είναι στο σημείο x_0. Η εκτίμηση γίνεται όχι μόνο στο σημείο x_0, αλλά και σε όλες τις n παρατηρήσεις, x_i. Μία γνωστή συνάρτηση στάθμισης, η οποία χρησιμοποιείται συχνά, είναι η τρικυβική συνάρτηση:

$$W(z) = \begin{cases} (1 - |z|^3)^3 & \text{για} \quad |z| < 1 \\ 0 & \text{για} \quad |z| \geq 1 \end{cases}.$$

Στο παρακάτω παράδειγμα χρησιμοποιείται το πλαίσιο δεδομένων `Prestige`, το οποίο βρίσκεται στη βιβλιοθήκη `car` και δίνει το βαθμό γοήτρου για τα διάφορα

επαγγέλματα στον Καναδά. Θα εξεταστεί η σχέση του βαθμού γοήτρου (prestige) με το εισόδημα (income). Για τη εφαρμογή της πολυωνυμικής παλινδρόμησης των τοπικά σταθμισμένων ελαχίστων τετραγώνων στην R χρησιμοποιείται η συνάρτηση lowess (local weighted scatter plot smoothing). Το όρισμα f δίνει το περίβλημα του εξομαλυντή, δηλαδή την αναλογία των σημείων στο γράφημα τα οποία επηρεάζουν την εξομάλυνση σε κάθε σημείο, ενώ το όρισμα iter δίνει τον αριθμό των επαναλήψεων που θα εκτελεστούν για τη διαδικασία εκτίμησης με σκοπό τη μείωση της βαρύτητας στα τελικά αποτελέσματα των απομακρυσμένων παρατηρήσεων. Το διάγραμμα διασπορών των δεδομένων μαζί με τη γραμμή εξομάλυνσης lowess παρουσιάζεται στο Σχήμα 18.1.

```
> library(car)
> attach(Prestige)
> plot(income,prestige,xlab="Average Income",ylab="Prestige")
> lines(lowess(income,prestige,f=0.5,iter=0),lwd=2)
```

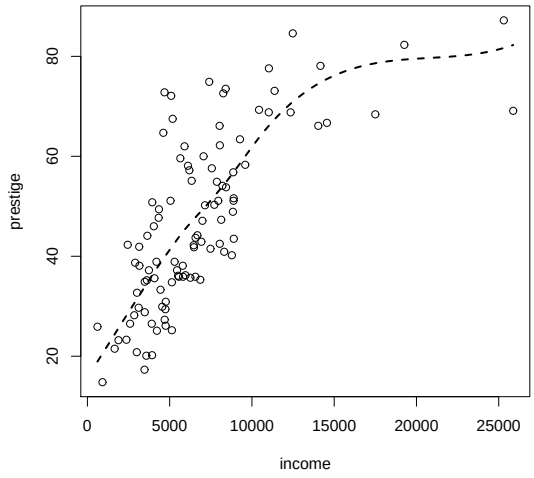


Σχήμα 18.1: Διάγραμμα διασποράς των δεδομένων μαζί με τη γραμμή εξομάλυνσης lowess.

Στην R υπάρχει επίσης η συνάρτηση loess η οποία χρησιμοποιείται για την πιο πάνω μέθοδο, η οποία έχει και περισσότερες δυνατότητες. Ακολουθεί ένα παράδειγμα για το πως χρησιμοποιείται για τα πιο πάνω δεδομένα. Ορίζοντας

degree=1 εφαρμόζεται η τοπικά γραμμική παλινδρόμηση. Το Σχήμα 18.2 δίνει το διάγραμμα διασπορών των δεδομένων με τη γραμμή εξομάλυνσης loess.

```
> mod.lo.inc<-loess(prestige~income,span=0.7,degree=1)
> mod.lo.inc
Call: loess(formula = prestige ~ income, span = 0.7, degree = 1)

Number of Observations: 102 Equivalent Number of Parameters: 3.85
Residual Standard Error: 11.13
> inc.100<-seq(min(income),max(income),len=100)
> pres<-predict(mod.lo.inc,data.frame(income=inc.100))
> plot(income,prestige)
> lines(inc.100,pres,lty=2,lwd=2)
```



Σχήμα 18.2: Διάγραμμα διασποράς των δεδομένων μαζί με τη γραμμή εξομάλυνσης lowess.

18.2 Εξομαλυντές Splines

Οι εξομαλυντές Splines είναι η λύση του προβλήματος απλής παλινδρόμησης, το οποίο επιζητά την εύρεση των συναρτήσεων $\hat{f}(x)$ με δύο συνεχείς παραγώγους, οι

οποίες ελαχιστοποιούν το άθροισμα τετραγώνων ποινής (penalized sum of squares),

$$SS^*(h) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + h\int_{x_{min}}^{x_{max}}[f''(x)]^2 dx,$$

όπου h είναι η παράμετρος της εξομάλυνσης, η οποία θεωρείται ανάλογη του πλάτους της γειτονιάς των τοπικά πολυωνυμικών εκτιμητών. Ο πρώτος όρος της εξίσωσης πιο πάνω είναι το άθροισμα τετραγώνων των υπολοίπων, ενώ ο δεύτερος όρος είναι η ποινή τραχύτητας (roughness penalty). Η ποινή αυτή είναι μεγάλη όταν η ολοκληρωτική δεύτερη παράγωγος της συνάρτησης παλινδρόμησης $f''(x)$ είναι μεγάλη, δηλαδή όταν η $f(x)$ αλλάζει γρήγορα κλίση. Όταν η σταθερά εξομάλυνσης h είναι ίση με 0, τότε η $\hat{f}(x)$ απλά παρεμβάλει τα δεδομένα. Αυτό είναι παρόμοιο με την εκτίμηση με τοπική παλινδρόμηση με περίβλημα ίσο με $1/n$. Αν το h όμως είναι αρκετά μεγάλο, τότε η $\hat{f}(x)$ θα επιλεχθεί έτσι ώστε η $\hat{f}''(x)$ είναι παντού 0, η οποίο ουσιαστικά είναι ισοδύναμη με μια γενική γραμμική εφαρμογή ελαχίστων τετραγώνων στα δεδομένα.

Το Σχήμα 18.3 παρουσιάζει στο ίδιο γράφημα την εξομάλυνση τοπικής πολυωνυμικής παλινδρόμησης `loess`, και την εξομάλυνση splines, η οποία γίνεται με την εντολή `smooth.spline` που βρίσκεται στη βιβλιοθήκη `splines`. Το γράφημα αυτό κατασκευάζεται όπως το γράφημα στο Σχήμα 18.2 προστέθοντας στο τέλος την εξομάλυνση splines με τη βοήθεια της εντολής `lines` όπως πιο κάτω:

```
> library(splines)
> lines(smooth.spline(income,prestige,df=3.85),lwd=2)
> legend("bottomright",c("loess","smoothing splines"),lty=c(2,1),lwd=c(2,2))
```

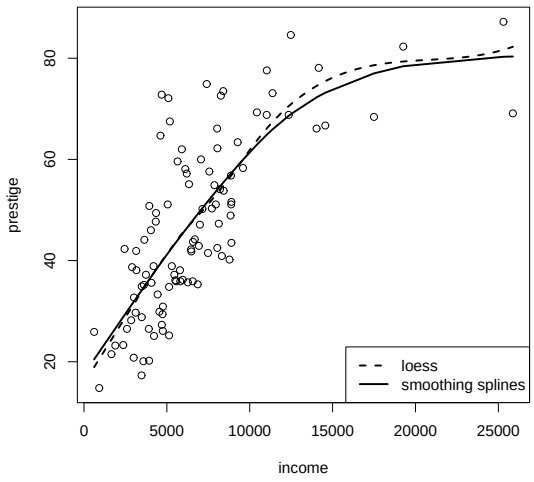
Οι βαθμοί ελευθερίας `df` τέθηκαν ίσοι με 3.85 για να συμφωνούν με τους βαθμούς ελευθερίας της εξομάλυνσης τοπικής πολυωνυμικής παλινδρόμησης.

18.3 Αθροιστική Απαραμετρική Παλινδρόμηση

Το μοντέλο της αθροιστικής μη παραμετρικής παλινδρόμησης δίνεται από

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + +f_k(x_{ik}) + \varepsilon_i,$$

όπου οι μερικές συναρτήσεις παλινδρόμησης f_j εφαρμόζονται χρησιμοποιώντας ένα εξομαλυντή απλής παλινδρόμησης, όπως για παράδειγμα η τοπικά πολυωνυμική παλινδρόμηση ή ο εξομαλυντής splines. Για την εφαρμογή της μεθόδου για την παλινδρόμηση του βαθμού γοήτρου συναρτήσει του εισοδήματος και της



Σχήμα 18.3: Διάγραμμα διασποράς των δεδομένων μαζί με τη γραμμή εξομάλυνσης lowess και smoothing splines.

εκπαίδευσης, χρησιμοποιείται η εντολή gam που βρίσκεται στη βιβλιοθήκη mgcv, όπως πιο κάτω:

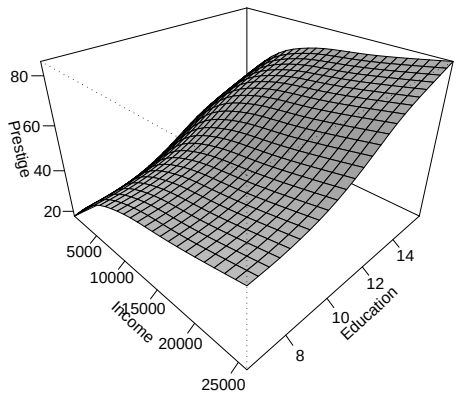
```
> library(mgcv)
> mod.gam<-gam(prestige~s(income)+s(education))
> mod.gam

Family: gaussian Link function: identity

Formula: prestige ~ s(income) + s(education)

Estimated degrees of freedom:
 3.117833 3.177297   total =  7.29513

GCV score:  52.1428
```

Η συνάρτηση s στην εντολή gam υποδεικνύει ότι κάθε όρος θα αναλυθεί χρησιμοποιώντας εξομαλυντή spline. Οι βαθμοί ελευθερίας υπολογίζονται με γενικευμένη σταυρωτή επαλήθευση. Σε αυτήν την περίπτωση, 3.1178 παράμετροι έχουν

χρησιμοποιηθεί για τον όρο `income`, και 3.1773 για τον όρο `education`. Οι βαθμοί ελευθερίας του μοντέλου είναι ίσοι με το άθροισμα τους συν 1, τη σταθερά της παλινδρόμησης.

Η επιφάνεια της αθροιστικής παλινδρόμησης δίνεται στο Σχήμα 18.4 και κατασκευάζεται όπως πιο κάτω:

```
> inc<-seq(min(income),max(income),len=25)
> ed<-seq(min(education),max(education),len=25)
> newdata<-expand.grid(income=inc,education=ed)
> fit.prestige<-matrix(predict(mod.gam,newdata),25,25)
> persp(inc,ed,fit.prestige,theta=45,phi=30,ticktype="detailed",
+ xlab="Income",ylab="Education",zlab="Prestige",expand=2/3,
+ shade=0.5)
```

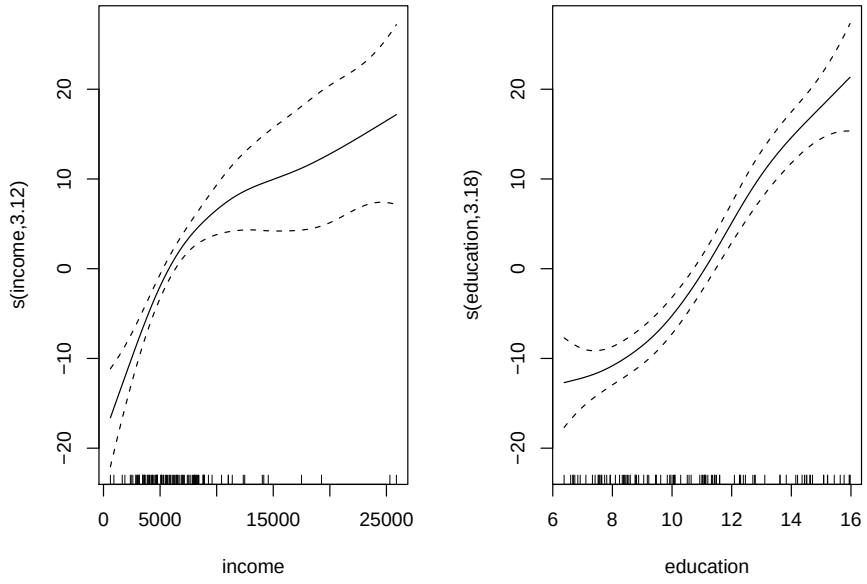


Σχήμα 18.4: Εκτιμώμενη επιφάνεια της αθροιστική απαραμετρικής παλινδρόμηση.

Για το λόγο ότι κομμάτια της επιφάνειας στην κατεύθυνση της μίας επεξηγηματικής μεταβλητής είναι παράλληλα, είναι αρκετό να γίνει το γράφημα κάθε μερικής παλινδρόμησης ξεχωριστά. Αυτή είναι και μια χρησιμότητα του μοντέλου αθροιστικής παλινδρόμησης, μειώνει το πολυδιάστατο πρόβλημα παλινδρόμησης σε μια σειρά από διδιάστατα γραφήματα των μερικών παλινδρομήσεων. Η σειρά

των γραφημάτων αυτών κατασκευάζεται στην R χρησιμοποιώντας το αντικείμενο gam ως όρισμα στην εντολή plot (Σχήμα 18.5).

```
> par(mfrow=c(1,2))
> plot(mod.gam)
```



Σχήμα 18.5: Γραφήματα των μερικών παλινδρομήσεων του μοντέλου αθροιστικής απαραμετρικής παλινδρόμηση.