Κεφάλαιο 8

Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση είναι ένα από τα πιο σημαντικά θέματα της Στατιστική θεωρείας. Στη συνέχεια αυτή η πολύ γνωστή μεθοδολογία θα αναπτυχθεί στην R μέσω των τύπων για τα μοντέλα.

8.1 Γραμμικά Μοντέλα στην R

Ο τύπος είναι μια έκφραση της R η οποία καθορίζει τη μορφή του μοντέλου με τις ανάλογες μεταβλητές. Για παράδειγμα, για να καθοριστεί ότι η Y είναι γραμμικός συνδυασμός δύο επεξηγηματικών μεταβλητών $X1$ και $X2$, χρησιμοποιείται ο ακόλουθος τύπος :

```
Y ~ X1 + X2.
```

Η περισπωμένη διαχωρίζει την εξαρτημένη μεταβλητή από τις επεξηγηματικές μεταβλητές. Με άλλα λόγια, εφαρμόζεται το μοντέλο

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2.$$

Ο τύπος *πάντα* υπονοεί την ύπαρξη του σταθερού όρου στο μοντέλο (β_0 στον πιο πάνω τύπο). Ωστόσο, είναι δυνατόν να αφαιρεθεί ο σταθερός όρος από το μοντέλο, προσθέτοντας στον τύπο του μοντέλου τον όρο -1 σαν επεξηγηματική μεταβλητή:

```
Y ~ -1 + X1 + X2.
```

Όταν ορίζονται κατηγορικές μεταβλητές,δηλαδή παράγοντες, σαν επεξηγηματικές μεταβλητές στα μοντέλα, η συνάρτηση μοντελοποίησης εφαρμόζει έναν σταθερό όρο για κάθε επίπεδο της μεταβλητής. Για παράδειγμα, για να κατασκευαστεί το γραμμικό μοντέλο με εξαρτημένη μεταβλητή το μισθό (`salary`) και επεξηγηματικές μεταβλητές την ηλικία (`age`), η οποία είναι συνεχής, και το φύλο (`gender`), η οποία είναι παράγοντας, ορίζεται όπως πιο κάτω:

```
salary ~ age + gender
```

Ωστόσο, διαφορετική παράμετρος εφαρμόζεται για κάθε ένα από τα δύο επίπεδα για το φύλο. Αυτό είναι ισοδύναμο με το να κατασκευαστεί το μοντέλο με δυο ψευδο-μεταβλητές, μία για *άρρεν* και μία για *θήλυ*. Συνεπώς δε χρειάζεται να οριστούν οι ψευδο-μεταβλητές στο μοντέλο.

Οι κύριες εκφράσεις για ορισμό γραμμικού μοντέλου είναι οι ακόλουθες

- `Y~X`: Γραμμικό μοντέλο του Y συναρτήσει του X
- `X1+X2`: Συμπεριλαμβάνει το $X1$ και το $X2$ στο μοντέλο
- `X1-X2`: Συμπεριλαμβάνει όλα από το $X1$ εκτός από αυτά που βρίσκονται στο $X2$ στο μοντέλο
- `X1:X2`: Συμπεριλαμβάνει την αλληλεπίδραση μεταξύ $X1$ και $X2$, $X1 : X2$, στο μοντέλο
- `X1*X2`: Όλο το μοντέλο $X1 + X2 + X1 : X2$

Οι επόμενες ενότητες εφαρμόζουν αυτές τις έννοιες σε μοντέλο πολλαπλής γραμμικής παλινδρόμισης.

8.2 Πολλαπλή Γραμμική Παλινδρόμηση

Το πλαίσιο δεδομένων `Trees` είναι ένα δείγμα δέντρων μαύρης κερασιάς. Στον ακόλουθο πίνακα παρουσιάζονται μερίκες από τις μετρήσεις για τη διάμετρο (σε ίντσες), το ύψος (σε πόδια) και τον όγκο (σε κυβικά πόδια). Η πλήρης συλλογή δεδομένων βρίσκεται στο παράρτημα αυτού του κεφαλαίου.

Ο σκοπός της συλλογής αυτών των δεδομένων ήταν για να βρεθεί ένας τρόπος πρόβλεψης του όγκου της ξυλείας των δέντρων από τις μετρήσεις για το ύψος και τη διάμετρό τους, χρησιμοποιώντας γραμμικό μοντέλο. Σε αυτήν την περίπτωση

Diameter	Height	Volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7

Πίνακας 8.1: Οι πρώτες έξι παρατηρήσεις από το σχετικό πλαίσιο δεδομένων.

η εξαρτημένη μεταβλητή είναι συνεχής και το αρχικό μοντέλο που θα εξεταστεί είναι το συνηθισμένο γραμμικό μοντέλο, με γενική μορφή

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

όπου Y είναι η εξαρτημένη μεταβλητή, $X_1, \ldots, X_p$ το σύνολο των επεξηγηματικών μεταβλητών, και ϵ το υπόλοιπο. Οι συντελεστές παλινδρόμησεις β_i εκτιμούνται με τη μέθοδο ελαχίστων τετραγώνων υποθέτωντας ότι το ϵ ακολουθεί την κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση σ^2. Για n παρατηρήσεις για την εξαρτημένη και τις επεξηγηματικές μεταβλητές, το μοντέλο μπορεί να γραφεί συνοπτικά ως

$$E(\mathbf{Y}) = \mathbf{X}\beta.$$

Η ανάλυση των γραμμικών μοντέλων στην R γίνεται με την εντολή `lm()` όπως παρουσιάζεται πιο κάτω:

```
> trees<-read.table("trees.txt")
> names(trees)<-c("Diameter","Height","Volume")
> trees.fit <- lm(Volume~ Diameter+Height, trees)
> trees.fit

Call:
lm(formula = Volume ~ Diameter + Height, data = trees)

Coefficients:
(Intercept)     Diameter       Height
   -57.9877       4.7082       0.3393
> summary(trees.fit)
```

```
Call:
lm(formula = Volume ~ Diameter + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
> anova(trees.fit)
Analysis of Variance Table

Response: Volume
          Df Sum Sq Mean Sq  F value  Pr(>F)
Diameter   1 7581.8  7581.8 503.1503 < 2e-16 ***
Height     1  102.4   102.4   6.7943 0.01449 *
Residuals 28  421.9    15.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> trees.res <- residuals(trees.fit)
> trees.prd <- predict(trees.fit)
```

Η εντολή summary χρησιμοποιείται για τον t έλεγχο για τους συντελεστές της παλινδρόμησης, με μηδενική υπόθεση $\beta_i = 0$. Τα αποτελέσματα μας οδηγούν στο συμπέρασμα ότι οι συντελεστές παλινδρόμησης για τη διάμετρο και το ύψος είναι σημαντικά διαφορετικοί από 0. Επίσης, δίνεται εκτίμηση για τη διακύμανση των υπολοίπων, όπως και ο συντελεστής μεταβλητότητας R^2 από τον οποίο συμπε-

ραίνεται ότι περίπου το 95% της διακύμανσης του όγκου των δέντρων εξηγείται από τις δυο επεξηγηματικές μεταβλητές. Τέλος, δίνεται η τιμή της στατιστικής συνάρτησης F για τον στατιστικό έλεγχο ο οποίος ελέγχει αν όλοι οι συντελεστές παλινδρόμησης είναι ταυτόχρονα ίσοι με 0, το οποίο και απορρίπτεται (γιατί;). Με την εντολή `anova` παρουσιάζεται ο πίνακας ανάλυσης διακύμανσης (ΑΝΑΔΙΑ), ενώ οι εντολές `residuals` και `predict` δίνουν τα υπόλοιπα και τις εκτιμήσεις του μοντέλου, αντίστοιχα.

Το επόμενο στάδιο στην ανάλυση είναι η ανάλυση των υπολοίπων του μοντέλου, δηλαδή της διαφοράς μεταξύ των αρχικών παρατηρήσεων και των εκτιμώμενων από το μοντέλο τιμών. Αυτή γίνεται κυρίως γραφικά, και οι πιο χρήσιμες γραφικές παραστάσεις είναι οι πιο κάτω:

1. Γραφική παράσταση των υπολοίπων συναρτήσει των επεξηγηματικών μεταβλητών του μοντέλου. Η παρουσία καμπυλόγραμμης σχέσης, για παράδειγμα, εισηγείται την πρόσθεση ενός όρου μεγαλύτερου βαθμού, ίσως τετραγωνικού, στο μοντέλο (Σχήμα 8.1).

2. Γραφική παράσταση των υπολοίπων συναρτήσει των εκτιμώμενων τιμών της εξαρτημένης μεταβλητής. Αν η διακύμανση της εξαρτημένης μεταβλητής φαίνεται να μεγαλώνει μαζί με την εκτιμώμενη τιμή, είναι δυνατό να χρειαστεί να γίνει μετασχηματισμός της εξαρτημένης μεταβλητής (Σχήμα 8.2).

3. QQ-γράφημα των υπολοίπων. Μετά την αφαίρεση όλης της συστηματικής διασποράς από τα δεδομένα, τα υπόλοιπα πρέπει να μοιάζουν με ένα δείγμα από την κανονική κατανομή. Είναι το γράφημα των ποσοστημορίων των υπολοίπων συναρτήσει των αναμενόμενων ποσοστημορίων από την κανονική κατανομή (Σχήμα 8.3).

Θα εργαστούμε με τα τυποποιημένα υπόλοιπα τα οποία ορίζονται από

$$r_i = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}$$

όπου h_{ii} τα διαγώνια στοιχεία του πίνακα

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}.$$

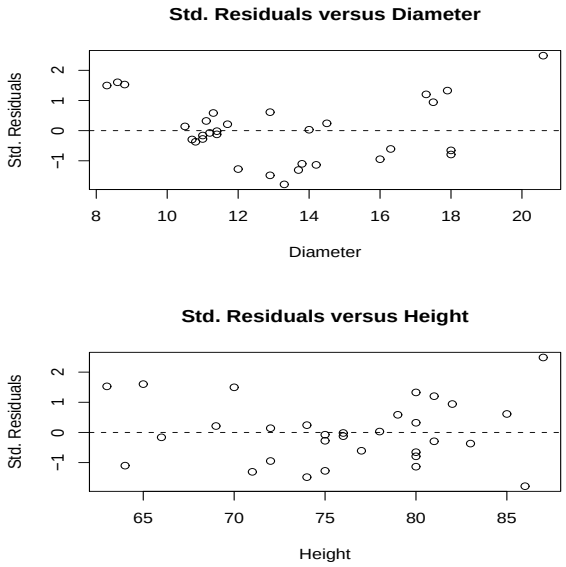
Στην R έχουμε ότι:

```
> s <- summary(trees.fit)$sigma
> h <- lm.influence(trees.fit)$hat
> trees.res <- trees.res/(s*sqrt(1-h))
```

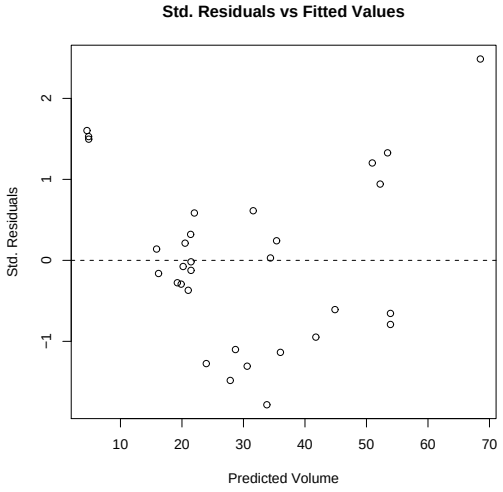
Η πρώτη εντολή εξάγει την εκτιμήτρια για το σ, η δεύτερη δίνει τα διαγώνια στοιχεία του πίνακα $\mathbf{H}$ και η τελευταία υπολογίζει τα τυποποιημένα υπόλοιπα. Ακολουθούν τα γραφήματα των υπολοίπων (Σχήματα 8.1 - 8.3) τα οποία κατασκευάζονται με τις πιο κάτω εντολές :

```
> par(mfrow=c(2,1))
> plot(trees[,"Diameter"], trees.res, xlab="Diameter",
+ ylab="Std. Residuals")
> abline(h=0, lty=2)
> title("Std. Residuals versus Diameter")
> plot(trees[,"Height"], trees.res, xlab="Height",
+ ylab="Std. Residuals")
> abline(h=0, lty=2)
> title("Std. Residuals versus Height")
> par(mfrow=c(1,1))
> plot(trees.prd, trees.res, xlab="Predicted Volume",
+ ylab="Std. Residuals")
> abline(h=0,lty=2)
> title("Std. Residuals vs Fitted Values")
> qqnorm(trees.res,ylab="Std. Residuals",
+ main="Normal Plot of Std. Residuals")
> qqline(trees.res)
```

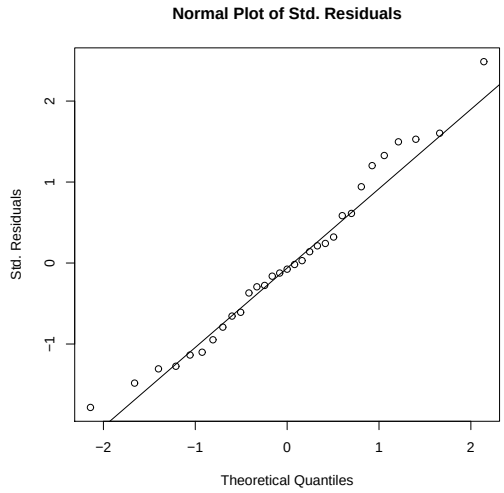
Τα γραφήματα των υπολοίπων συναρτήσει τη διαμέτρου των δέντρων, αλλά και τις εκτιμώμενες τιμές, δείχνουν ότι ένας τετραγωνικός όρος θα μπορούσε να προστεθεί στο μοντέλο. Η ερμηνεία του QQ-γραφήματος συχνά δεν μπορεί να είναι ξεκάθαρη, ειδικά στις περιπτώσεις με μικρά δείγματα. Ωστόσο, εξετάζοντας το φαίνεται ότι τα υπόλοιπα έχουν μικρή απόκλιση από την κανονική.



Σχήμα 8.1: Τυποποιημένα υπόλοιπα συναρτήσει των επεξηγηματικών μεταβλητών.



Σχήμα 8.2: Τυποποιημένα υπόλοιπα συναρτήσει των εκτιμώμενων τιμών της εξαρτημένης μεταβλητης.

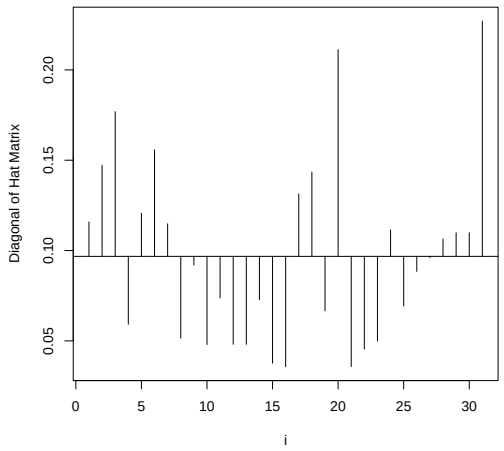


Σχήμα 8.3: QQ-γράφημα των τυποποιημένων υπολοίπων.

Ο πίνακας $\mathbf{H}$ είναι επίσης πολύ βοηθητικός στην αναγνώριση παράξενων ή ασυνήθιστων σημείων των δεδομένων, τα οποία συχνά έχουν μεγάλη επίδραση στην γραμμική παλινδρόμηση. Τέτοια σημεία αναγνωρίζονται από τις σχετικά μεγάλες τιμές στην αντίστοιχη θέση στη διαγώνιο του $\mathbf{H}$. Η μεγαλύτερη τιμή σε οποιοδήποτε στοιχείο της διαγωνίου είναι το 1. Τεχνικά αυτά τα σημεία φαίνονται να έχουν μεγάλη επιρροή (leverage) (Σχήμα 8.4).

```
> h <- lm.influence(trees.fit)$hat
> h
         1          2          3          4          5          6
0.11582883 0.14720958 0.17686186 0.05919131 0.12066468 0.15575111
         7          8          9         10         11         12
0.11480262 0.05148096 0.09200658 0.04797237 0.07382512 0.04809206
        13         14         15         16         17         18
0.04809206 0.07275901 0.03764563 0.03566543 0.13130916 0.14346152
        19         20         21         22         23         24
0.06665975 0.21123665 0.03580935 0.04541796 0.04994875 0.11142518
        25         26         27         28         29         30
0.06930648 0.08841762 0.09603041 0.10641665 0.10982638 0.10982638
        31
 0.22705852
```

```
> plot(1:31,h, type="n", xlab="i", ylab="Diagonal of Hat Matrix")
> abline(h=mean(h))
> segments(1:31, h, 1:31, mean(h))
```



Σχήμα 8.4: Γράφημα επιρροής.

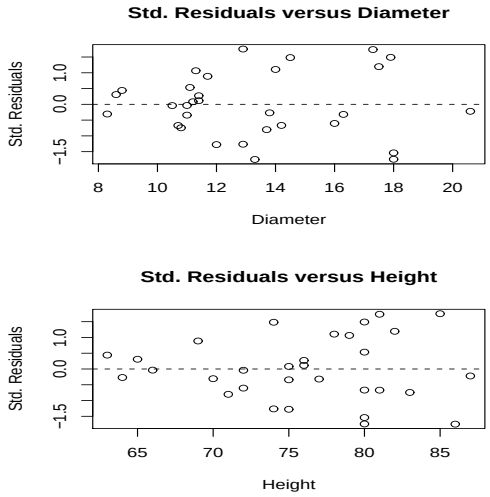
Εδώ δε φαίνεται να υπάρχουν οποιαδήποτε προβληματικά σημεία τα οποία είναι δυνατό να επηρεάσουν υπερβολικά τη διαδικασία εκτίμησης. Οι τιμές της επιρροής είναι σχετικά μικρές.

Επιστρέφοντας τώρα στην ένδειξη από τα γραφήματα των υπολοίπων, θα μελετηθεί ένα νέο μοντέλο το οποίο περιέχει τον τετραγωνικό όρο για τη διάμετρο.

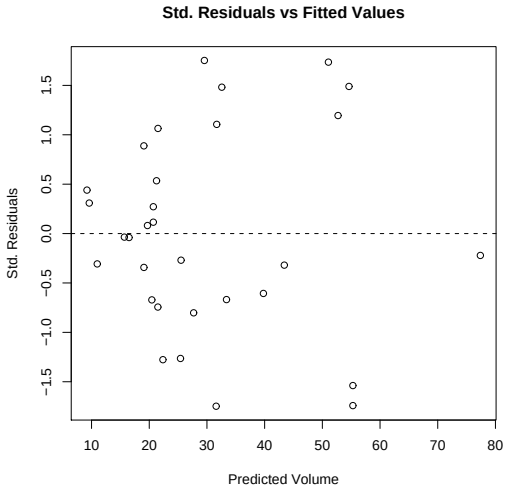
```
> trees1.fit <- lm(Volume~Diameter+I(Diameter*Diameter)+Height,
+ trees)
> trees1.fit

Call:
lm(formula=Volume~Diameter+I(Diameter*Diameter)+ Height,data=trees)

Coefficients:
          (Intercept)              Diameter  I(Diameter * Diameter)
              -9.9204               -2.8851                   0.2686
               Height
```

```
               0.3764

> summary(trees1.fit)

Call:
lm(formula = Volume ~ Diameter + I(Diameter * Diameter) + Height,
    data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -9.92041   10.07911  -0.984 0.333729
Diameter              -2.88508    1.30985  -2.203 0.036343 *
I(Diameter * Diameter)  0.26862    0.04590   5.852 3.13e-06 ***
Height                 0.37639    0.08823   4.266 0.000218 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,     Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16
> trees1.res<-residuals(trees1.fit)
> trees1.prd<-predict(trees1.fit)
```

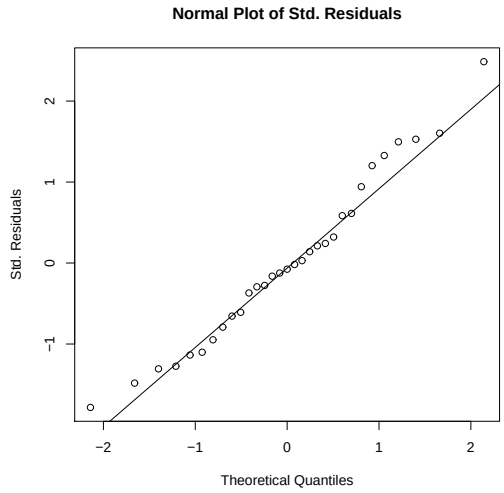
Τα γραφήματα των υπολοίπων κατασκευάζονται όπως προηγουμένως και παρουσιάζονται πιο κάτω (Σχήματα 8.5-8.7). Φαίνεται ότι στο γράφημα των νέων τυποποιημένων υπολοίπων συναρτήσει της διαμέτρου δεν υπάρχει πλέον η συστηματικότητα που υπήρχε πριν. Τα υπόλοιπα γραφήματα κρίνονται ικανοποιητικά.



Σχήμα 8.5: Τυποποιημένα υπόλοιπα δευτεροβάθμιου μοντέλου συναρτήσει των επεξηγηματικών μεταβλητών.



Σχήμα 8.6: Τυποποιημένα υπόλοιπα δευτεροβάθμιου μοντέλου συναρτήσει των εκτιμώμενων τιμών της εξαρτημένης μεταβλητης.



Σχήμα 8.7: QQ-γράφημα των τυποποιημένων υπολοίπων δευτεροβάθμιου μοντέλου.

Παρόλο που τα αποτελέσματα της πολλαπλής γραμμικής παλινδρόμησης υποδεικνύουν ότι οι συντελεστές παλινδρόμησης και για τη διάμετρο και για το ύψος είναι σημαντικά διαφορετικοί από μηδέν, πολύ συχνά είναι χρήσιμο να ερευνηθεί ένας αριθμός μοντέλων σε μια προσπάθεια να βρεθεί το πιο απλό μοντέλο που εφαρμόζει καλύτερα τα δεδομένα. Ουσιαστικά, η διαδικασία επιλογής μοντέλου περιλαμβάνει την πρόσθεση ή την αφαίρεση όρων από ένα προυπάρχων μοντέλο και τον υπολογισμό της επίδρασης της αλλαγής. Υπολογισμός αυτός γίνεται με τη βοήθεια του κριτηρίου πληροφορίας του Akaike (AIC),το οποίο είναι ένα μέτρο της καλής εφαρμογής των δεδομένων από το μοντέλο. Όσο πιο μικρό το AIC τόσο καλύτερο είναι το μοντέλο.

Αρχικά, από το μοντέλο το οποίο περιέχει τη διάμετρο και το ύψος αφαιρείται μια μια μεταβλητή και υπολογίζεται η αλλαγή με το AIC. Όπως φαίνεται, αν αφαιρεθεί οποιαδήποτε από τις μεταβλητές το AIC μεγαλώνει και άρα το αρχικό μοντέλο εφαρμόζει καλύτερα τα δεδομένα. Επίσης διαφαίνεται και η σημαντικότητα της μεταβλητής της διαμέτρου στο μοντέλου αφού αν αφαιρεθεί μεγαλώνει σημαντικά το AIC.

```
> attach(trees)
> trees.drop1 <- drop1(trees.fit)
> trees.drop1
```

```
Single term deletions

Model:
Volume ~ Diameter + Height
         Df Sum of Sq    RSS    AIC
<none>                 421.9   86.9
Diameter  1    4783.0 5204.9  162.8
Height    1     102.4  524.3   91.7
```

Αντίθετα τώρα, ξεκινώντας από το μοντέλο που περιέχει μόνο τον σταθερό όρο, προσθέτουμε μια μια τις μεταβλητές.

```
> trees0.fit <- lm(Volume~1)
> trees.add1 <- add1(trees0.fit,~ Height+Diameter)
> trees.add1
Single term additions

Model:
Volume ~ 1
         Df Sum of Sq    RSS    AIC
<none>                8106.1  174.6
Height    1    2901.2 5204.9  162.8
Diameter  1    7581.8  524.3   91.7
```

Από τα αποτελέσματα συμπεραίνεται και πάλι η σημαντικότητα της ύπαρξης και των δυο μεταβλητών στο μοντέλο αφού το AIC γίνεται μικρότερο με την πρόσθεση τους. Εφαρμόζοντας τώρα τις εντολές `drop1` και `add1` στο δευτεροβάθμιο μοντέλο παρατηρείται ότι όλοι οι παράγοντες είναι σημαντικοί για το μοντέλο.

```
> trees1.drop1 <- drop1(trees1.fit)
> trees1.drop1
Single term deletions

Model:
Volume ~ Diameter + I(Diameter * Diameter) + Height
                       Df Sum of Sq    RSS    AIC
<none>                                186.01  63.55
Diameter                1     33.42 219.44  66.67
I(Diameter * Diameter)  1    235.91 421.92  86.94
```

```
Height                   1    125.37 311.38  77.52
>
> trees10.fit <- lm(Volume~1)
> trees1.add1 <- add1(trees10.fit,~ Height+Diameter+I(Diameter*Diameter))
> trees1.add1
Single term additions

Model:
Volume ~ 1
                       Df Sum of Sq    RSS    AIC
<none>                              8106.1  174.6
Height                  1    2901.2 5204.9  162.8
Diameter                1    7581.8  524.3   91.7
I(Diameter * Diameter)  1    7776.8  329.3   77.3
```

Παράρτημα

Τα δεδομένα που χρησιμοποιούνται σε αυτό το κεφάλαιο για εφαρμογή της πολλαπλής παλινδρόμησης.

```
> trees
   Diameter Height Volume
1       8.3     70   10.3
2       8.6     65   10.3
3       8.8     63   10.2
4      10.5     72   16.4
5      10.7     81   18.8
6      10.8     83   19.7
7      11.0     66   15.6
8      11.0     75   18.2
9      11.1     80   22.6
10     11.2     75   19.9
11     11.3     79   24.2
12     11.4     76   21.0
13     11.4     76   21.4
14     11.7     69   21.3
15     12.0     75   19.1
16     12.9     74   22.2
17     12.9     85   33.8
18     13.3     86   27.4
19     13.7     71   25.7
20     13.8     64   24.9
21     14.0     78   34.5
22     14.2     80   31.7
23     14.5     74   36.3
24     16.0     72   38.3
25     16.3     77   42.6
26     17.3     81   55.4
27     17.5     82   55.7
28     17.9     80   58.3
29     18.0     80   51.5
30     18.0     80   51.0
```

```
31      20.6      87   77.0
```