Παλινδρόμηση

Περιεχόμενα:

Απλή παλινδρόμηση

Στην πράξη συχνά θέλουμε να εξετάσουμε περισσότερες από μία μεταβλητές. Συνήθως θέλουμε να δούμε πως συσχετίζεται μία μεταβλητή με μία άλλη.

Η ανάλυση παλινδρόμησης χρησιμοποιείται για την επεξήγηση ή μοντελοποίηση της σχέσης μεταξύ μίας μεταβλητής y, που ονομάζεται *μεταβλητή απόκρισης*, *εξαγόμενη* ή *εξαρτημένη* μεταβλητή, και ενός ή περισσότερους *εκτιμητές* , *εισαγομένων* , *ανεξαρτήτων* , ή *επεξηγηματικών* μεταβλητών $x_1, x_2, \ldots, x_n$. Αν $n = 1$, τότε ονομάζεται απλή παλινδρόμηση· διαφορετικά, αν $n > 1$ ονομάζεται πολλαπλή παλινδρόμηση, ή πολυμεταβλητή παλινδρόμηση. Στην περίπτωση που υπάρχουν περισσότερα από ένα y, τότε ονομάζεται πολλαπλή πολυμεταβλητή ανάλυση.

Η παλινδρόμηση έχει διάφορους στόχους μεταξύ:

- Πρόβλεψη μελλοντικών παρατηρήσεων·

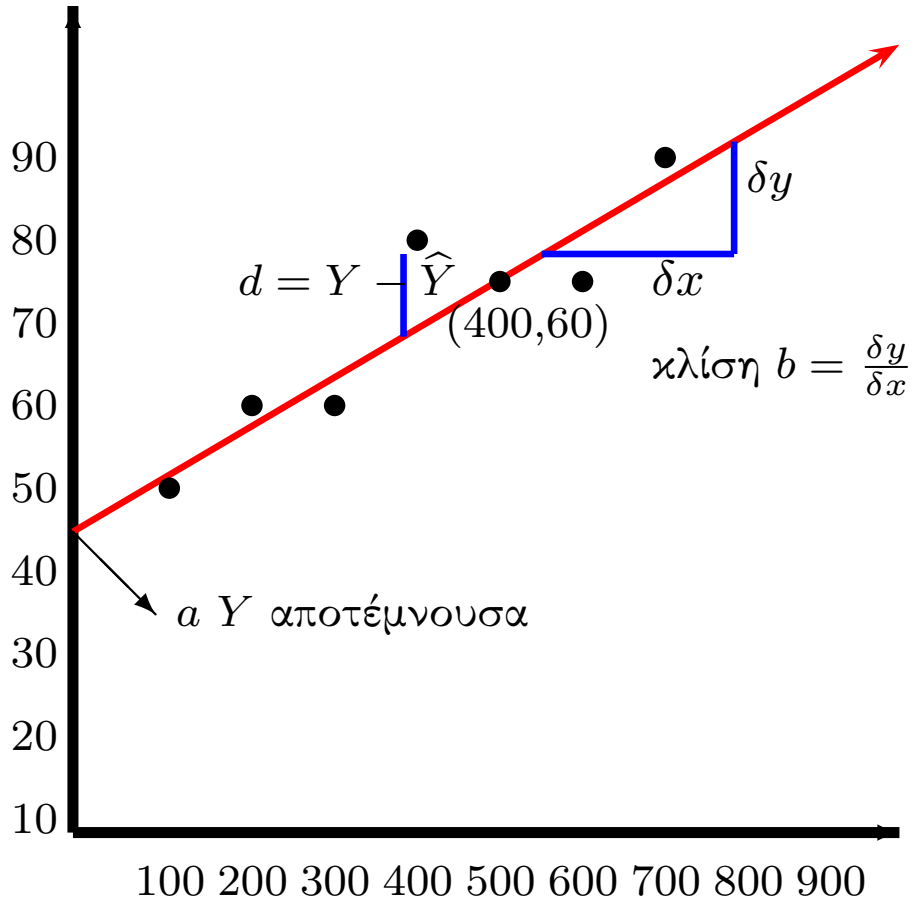
- Αξιολόγηση της επίδρασης ή της σχέσης μεταξύ επεξηγηματικών μεταβλητών και εξαγόμενης μεταβλητής·
- Γενική περιγραφή της δομής των δεδομένων.

Παράδειγμα

Για μια συγκεκριμένη κατηγορία ασφαλειών θέλουμε να ερευνήσουμε πως τα ασφάλιστρα συσχετίζονται με τις απαιτήσεις. Θέτουμε X τις απαιτήσεις και Y τα ασφάλιστρα. Το σύνολο επτά διαφορετικών επιπέδων δίνεται στον πιο κάτω πίνακα:

X	100	200	300	400	500	600	700
Y	40	50	50	70	65	65	80

- Αναπαραστηστε τα σημεία και διαγράψετε με το μάτι μια ευθεία που να ενώνει τα σημεία.


90
80
70
60
50
40
30
20
10
100 200 300 400 500 600 700 800 900
$d = Y - \widehat{Y}$
(400,60)
δy
δx
κλίση $b = \frac{\delta y}{\delta x}$
a Y αποτέμνουσα

Μέθοδος ελαχίστων τετραγώνων

Σκοπός είναι η ανάθεση μιας γραμμής της μορφής

$$\widehat{Y} = a + bX.$$

Έτσι, πρέπει να βρούμε τον τύπο για τον υπολογισμό της κλίσης b και της αποτέμνουσας a. Αυτός ο τύπος προέρχεται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων, δηλαδή,

$$\text{minimize} \sum d^2 = \sum (Y - \widehat{Y})^2.$$

Το πιο πάνω ονομάζεται κριτήριο *Κανονικών Ελαχίστων Τετραγώνων (Ordinary Least Squares)* (OLS) που οδηγεί στην επιλογή μια μοναδικής γραμμής που ονομάζεται γραμμή OLS.

Η κλίση OLS b υπολογίζεται από τον τύπο

$$\boxed{b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum xy}{\sum x^2},}$$

όπου $x = X - \bar{X}$, $y = Y - \bar{Y}$ και $x^2 = (X - \bar{X})^2$.

Η αποτέμνουσα a εξάγεται από

$$\boxed{a = \bar{Y} - b\bar{X}.}$$

Σημειώστε ότι η γραμμή ελαχίστων-τετραγώνων περνά από το σημείο $(\bar{X}, \bar{Y})$.

Παράδειγμα

Χρησιμοποιώντας τις τιμές του προηγούμενου παραδείγματος έχουμε:

$$\bar{X} = 400, \quad \text{και} \quad \bar{Y} = 60.$$

Επιπλέον , $x = X - \bar{X}$, $y = Y - \bar{Y}$, xy και x^2 υπολογίζονται από τον πίνακα:

x	-300	-200	-100	0	100	200	300
y	-20	-10	-10	10	5	5	20
$xy/1000$	6	2	1	0	0.5	1	6
$x^2/1000$	90	40	10	0	10	40	90

Σημειώστε ότι $\sum xy = 16500$ και $\sum x^2 = 280000$. Έτσι,

$$b = \sum xy / \sum x^2 = 0.059 \quad \text{και} \quad a = \bar{Y} - b\bar{X} = 36.4.$$

Έτσι, η γραμμή OLS δίνεται από:

$$\widehat{Y} = 36.4 + 0.059X.$$

Επομένως, αν $X = 400$, τότε το προβλεπόμενο

ασφάλιστρο $\hat{Y}$ δίνεται από

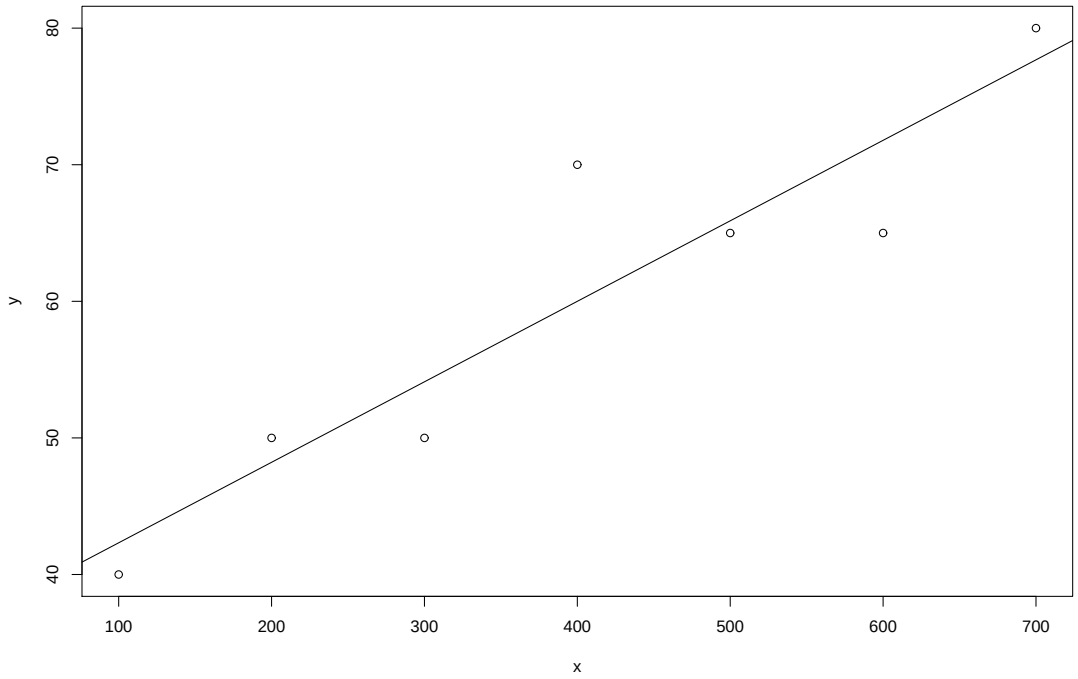
$$\widehat{Y} = 36.4 + 0.059 \times 400 = 60.$$

Η απόκλιση d των πραγματικών τιμών του Y από την προβλεπούσα τιμή $\widehat{Y}$ δίνεται από $d = Y - \widehat{Y}$.

Υπολογισμός παραγόντων

```
> x=c(100, 200, 300, 400,500, 600, 700);
> y=c(40, 50, 50, 70,65, 65, 80);
Residuals:
  1     2     3     4     5     6     7
-2.32  1.79 -4.11 10.00 -0.89 -6.79  2.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.038   7.231  0.00079
x            0.059    0.011   5.231  0.00338
---
Residual standard error: 5.961 on 5 df
Multiple R-Squared: 0.85, Adjusted R-squared: 0.81
F-statistic: 27.36 on 1 and 5 DF, p-value: 0.0034
```



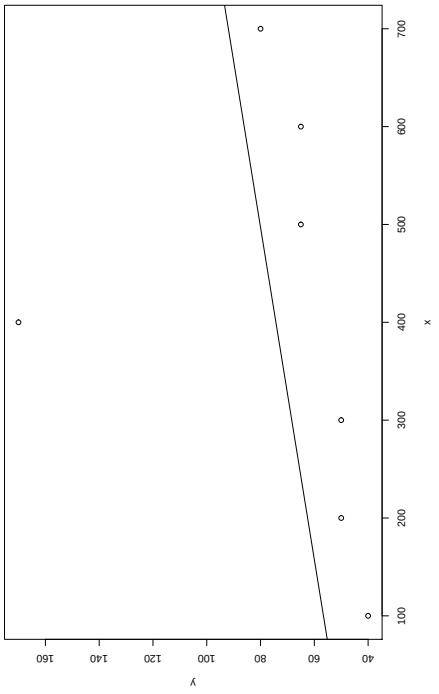
Θεωρήστε τις εναλλακτικές παρατηρήσεις $(400, 70)$ με $(400, 170)$ και $(400, 7)$. Πώς διαμορφώνονται οι OLS εκτιμητές;

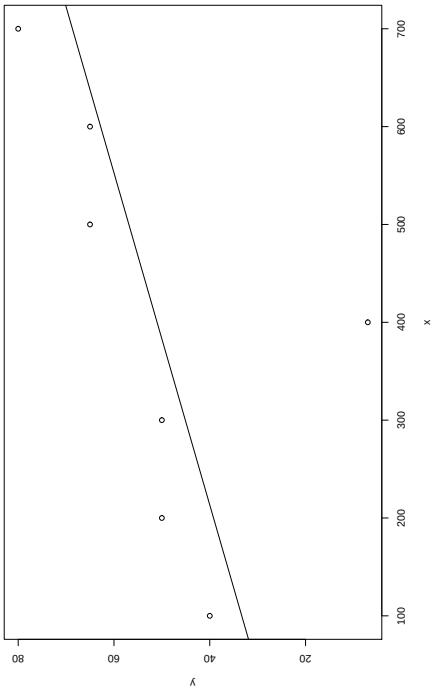
Η μεταβολή από: $(400, 70)$ σε $(400, 170)$

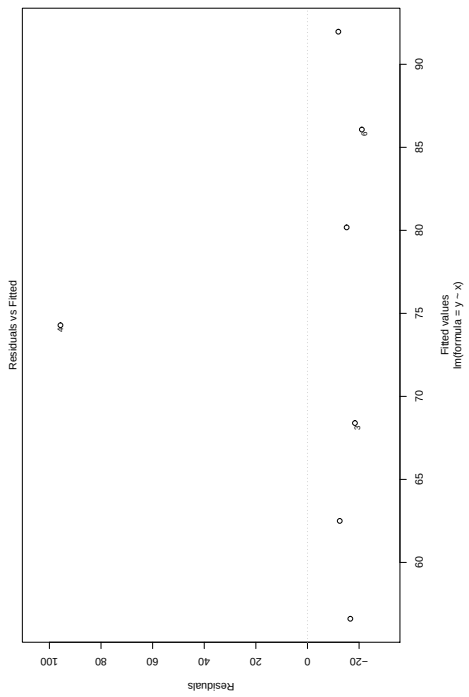
```
> y=c(40, 50, 50, 170,65, 65, 80);
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.714   39.19   1.29    0.252
x            0.059    0.09   0.67    0.531
-----
Residual standard error: 46.37 on 5 df
Multiple R-Squared: 0.08, Adjusted R-squared: -0.10
F-statistic: 0.4523 on 1 and 5 DF,  p-value: 0.53
```

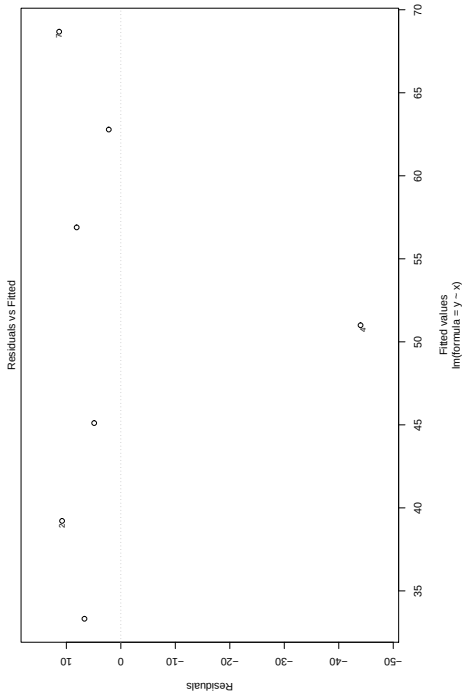
Η μεταβολή από: $(400, 70)$ σε $(400, 7)$

```
> y=c(40, 50, 50, 7,65, 65, 80);
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.429   18.20   1.51    0.192
x            0.059    0.04   1.45    0.207
------
Residual standard error: 21.54 on 5 df
Multiple R-Squared: 0.30, Adjusted R-squared: 0.15
F-statistic: 2.096 on 1 and 5 DF,  p-value: 0.21
```







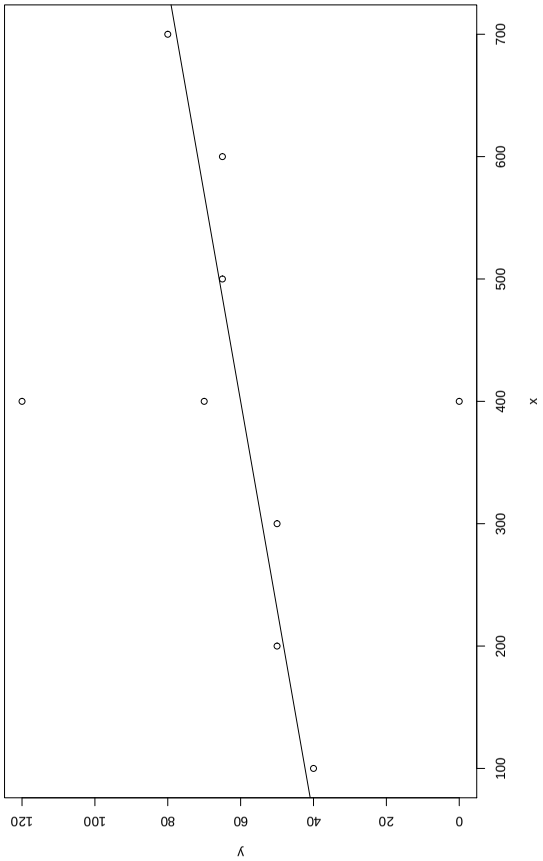


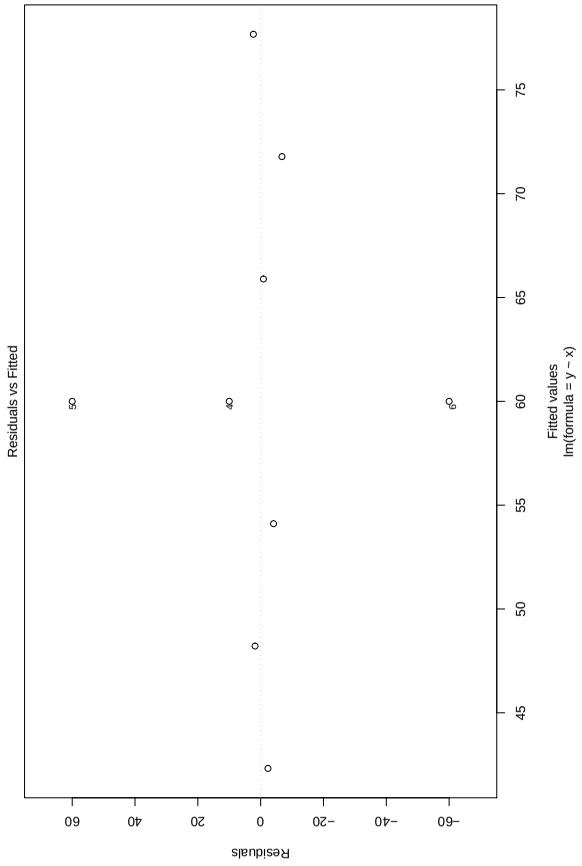
Οι παρατηρήσεις $(400, 170)$ και $(400, 7)$ είναι ανώμαλες, αλλά αφού είναι κοντά στο μέσο της επεξηγηματικής μεταβλητής, δεν επηρεάζεται η εκτιμώμενη κλίση.

Θεωρήστε την προσθήκη των παρατηρήσεων $(400, 120)$ και $(400, 0)$.

```
> x=c(100, 200, 300, 400,400,400,500, 600, 700);
> y=c(40, 50, 50, 70,120,0,65, 65, 80);
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.429   26.82   1.36    0.217
x            0.059    0.06   0.96    0.369
--------
Residual standard error: 32.46 on 7 df
```

```
Multiple R-Squared: 0.12, Adjusted R-squared: -0.01
F-statistic: 0.92 on 1 and 7 DF,  p-value: 0.37
```





Για την προσαρμογή των δεδομένων στο μοντέλο παλινδρόμησης γίνονται οι ακόλουθες υποθέσεις: Οι τυχαίες μεταβλητές απόκρισης $Y_1, \ldots, Y_n$ είναι ανεξάρτητες, με μέσο $a + bX_i$ και διακύμανση σ^2. Ωστόσο συχνά παρουσιάζουμε το μοντέλο στην πιο κάτω μορφή

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

όπου ε_i (ονομάζονται σφάλματα) υποδηλώνει την απόκλιση του Y από την αναμενόμενη τιμή. Σε αυτή την περίπτωση οι υποθέσεις είναι οι ακόλουθες: Τα σφάλματα $\varepsilon_1, \ldots, \varepsilon_n$ είναι ανεξάρτητα με μέσο 0 και διακύμανση σ^2.

Δειγματοληπτική μεταβλητότητα (Sampling variability

Θέλουμε να διερευνήσουμε πόσο ΄κοντά΄ η εκτιμημένη γραμμή (ευθεία) βρίσκεται στην `σωστή` πληθυσμιακή γραμμή. Συγκεκριμένα, πως κατανέμεται η εκτιμημόμενη κλίση b γύρω από το β.

Κανόνας κανονικής προσέγγισης παλινδρόμησης

Η κλίση της εκτίμησης b είναι προσεγγιστικά κανονικά κατανεμημένη με μέσο β και διακύμανση $\sigma^2/\sum x^2$. Έτσι,

$$\boxed{b \sim N(\beta,\, \sigma^2/\Sigma x^2).}$$

Σημειώστε ότι $\sum x^2 = \sum (X - \bar{X})^2 = n\, S_x^2$, όπου S_x^2 είναι η διακύμανση της μεταβλητής X. Συνεπώς,

$$\boxed{b \sim N\Big(\beta,\, \frac{\sigma^2}{nS_x^2}\Big).}$$

Η απόκλιση του b από το β αντιπροσωπεύει το εκτιμωμένο σφάλμα και ονομάζεται *Τυπικό Σφάλμα (Standard Error)* (SE). Το SE του b δίνεται από

$$\text{SE} = \frac{\sigma}{\sqrt{\Sigma x^2}} = \frac{\sigma}{\sqrt{n}}\ \frac{1}{S_x}.$$

Το πιο πάνω δείχνει ότι υπάρχουν τρεις τρόποι που μπορεί να μειωθεί το SE έτσι ώστε να έχουμε πιο ακριβείς εκτιμήσεις b:

1. Μειώνοντας το σ η συμφυής μεταβλητότητα των Y παρατηρήσεων.

2. Αυξάνοντας το δείγμα n.

3. Αυξάνοντας το S_x, την απόκλιση των τιμών X όπως ορίζονται από το πείραμα (έρευνα).

Θεωρήστε τη 'σωστή' σχέση:

$$y = a + bx,$$

όπου $a = 3.0$ και $b = 5$. Σκοπός είναι να εκτιμήσουμε τη σχέση που περιέχει τα σφάλματα ε.

Υποθέστε ότι δημιουργούμε τυχαίες τιμές για το x και το σφάλμα ε το οποίο είναι κανονικά κατανεμημένο με μέσο μηδέν και τυπική απόκλιση σ. Το y παράγεται από

$$y = a + bx + \varepsilon.$$

Οι εντολές για την πιο πάνω δημιουργία στο στατιστικό πακέττο R είναι οι ακόλουθες:

```
n <- 100
x <- runif(n, min=-200, max=200)
a <- 3.0
```

```
b <- 5
e <- rnorm(n, sd=1.5)
y <- a + b * x  + e
g <- lm(y~x)
summary(g)
```

Πώς οι εκτιμητές επηρεάζονται από το σφάλμα ε και τις τιμές του x.

1. Θέτουμε $\sigma = 1.0$ και επιλέγουμε τυχαία τιμές για το x από το όριο -5 μέχρι 5. Εάν το δείγμα είναι 10, τότε οι OLS εκτιμητές για το a και το b είναι $\hat{a} = 3.21$ και $\hat{\beta} = 4.78$. Αν το δείγμα αυξηθεί σε $n = 10000$, τότε οι εκτιμητές είναι $\hat{a} = 2.99$ και $\hat{\beta} = 5.00$.

2. Θεωρήστε την περίπτωση όπου οι τιμές του x επιλέγονται τυχαία από το όριο 5 μέχρι 5.2 και $\sigma = 1.0$. Με μέγεθος δείγματος $n = 10$ οι εκτιμητές είναι $\hat{a} = -43.36$ και $\hat{\beta} = 14.00$. Για $n = 10000$ βρίσκουμε ότι $\hat{a} = 2.34$ και $\hat{\beta} = 5.13$.

3. Στην τελευταία περίπτωση αν $\sigma = 0.01$ τότε για $n = 10$ $\hat{a} = 3.05$ και $\hat{\beta} = 4.99$, καθώς για

$n = 10000$ $\hat{a} = 2.99$ και $\hat{\beta} = 5.00$.

4. Αυξάνοντας το όριο των τιμών του x η περιοχή από -200 μέχρι 200 και $\sigma = 1.5$ δίνει $\hat{a} = 3.53$ και $\hat{\beta} = 4.99$ όταν $n = 10$. Για $n = 100$ οι εκτιμητές είναι $\hat{a} = 2.98$ και $\hat{\beta} = 5.00$

Η διακύμανση των Y παρατηρήσεων σ^2 είναι γενικά άγνωστη και πρέπει να εκτιμηθεί. Τα υπόλοιπα χρησιμοποιούνται για την εξαγωγη των εκτιμητών S^2 του σ^2. Έτσι,

$$\boxed{S^2 = \frac{1}{n-2}\sum (Y - \widehat{Y})^2.}$$

Σημειώστε ότι $(n-2)$ είναι οι βαθμοί ελευθερίας και $\sum(Y - \widehat{Y})$ είναι το επονομαζόμενο *Άθροιμα των τετραγώνων των σφαλμάτων*, *sum of squares of errors* (ΣΣΕ). Έτσι, $S^2 = \mathrm{SSE}/n - 2$, ο οποίος είναι αμερόληπτος εκτιμητής του σ^2. Επομένως, η εκτιμώμενη διακύμανση της κλίσης b δίνεται από $S^2/\sum x^2$.

Επιπλέον, το 95% διάστημα εμπιστοσύνη για το β

δίνεται από:

$$\boxed{\beta = b \pm T_{2.5\%}^{(n-2)} \frac{S}{\sqrt{\Sigma_{x^2}}}.}$$

Παράδειγμα

Από προηγούμενο παράδειγμα
$\widehat{Y} = 36.4 + 0.059 \times 400 = 60.$ Άρα λοιπόν:

$\widehat{Y}$	42.3	48.2	54.1	60.0	65.9	71.8	77.7
$Y - \widehat{Y}$	-2.3	1.8	-4.1	10.0	-0.9	-6.8	2.3
$(Y - \widehat{Y})^2$	5.29	3.24	16.81	100	0.81	46.24	5.29

Σημειώστε ότι SSE $= \sum (Y - \widehat{Y})^2 = 177.68$ και

$$S^2 = \frac{SSE}{n-2} = \frac{177.68}{5} = 35.5$$

$$\frac{S}{\sqrt{\Sigma x^2}} = \sqrt{\frac{35.5}{280000}} = 0.0113$$

$$T_{2.5\%}^{(n-2)} = T_{2.5\%}^{(5)} = 2.571.$$

Από το πιο πάνω παίρνουμε:

$$\begin{aligned} \beta &= b \pm T_{2.5\%}^{(n-2)} \frac{S}{\sqrt{\Sigma x^2}}. \\ &= 0.059 \pm 2.571 \times 0.0113 \\ &= 0.059 \pm 0.29, \end{aligned}$$

ή

$$0.030 < \beta < 0.088.$$

Η υπόθεση ότι X (απαιτήσεις) και Y (ασφάλιστρα) είναι ασυσχέτιστα μπορεί να εκφραστεί μαθηματικά ως $\beta = 0$. Εντούτοις σε 5% επιπέδου σφάλματος σημειώνουμε ότι το μηδέν δεν περιέχεται στο 95% διάστημα εμπιστοσύνης.

Επομένως, *σε 5% επίπεδο σφάλματος απορρίπτουμε την υπόθεση ότι τα ασφάλιστρα είναι ασυσχέτιστα με τις απαιτήσεις.*

P-value

Κάθε στατιστικός έλεγχος σχετίζεται με μια μηδενική υπόθεση, που συμβολίζεται H_0. Οι μηδενικές υποθέσεις δηλώνου ότι 'δεν υπάρχει επιδραση' ή 'δεν υπάρχει διαφορά'. Η p-value είναι η πιθανότητα το δείγμα να προέρχεται από τον πληθυσμό που εξετάζεται (ή ότι ένα πιο απίθανο δείγμα μπορεί να επιλεγεί) βάση της υπόθεσης ότι η μηδενική υπόθεση είναι αληθής. P-value ίσο με 0.05, για παράδειγμα, υποδηλώνει ότι θα υπήρχε μόνο 5% πιθανότητα επιλογής του 'σωστού' δείγματος εάν η μηδενική υπόθεση ήταν αληθής.

Ένα p-value κοντά στο μηδέν υποδηλώνει ότι η μηδενική υπόθεση είναι λανθασμένη και πολύ πιθανό να υπάρχει διαφορά. Μεγάλα p-values πιο κοντά στο 1 υπονοούν ότι δεν υπάρχει εμφανής διαφορά για το μέγεθος τους δείγματος που χρησιμοποιείται. P-value ίσο με 0.05 είναι το κατώτατο όριο που χρησιμοποιείται συνήθως στην πράξη για την αξιολόγηση της μηδενικής υπόθεσης. Σε πιο

κρίσιμους τομείς (υγεία, κτλ.) πιο αυστηρό, μικρότερο p-value μπορεί να εφαρμοστεί.

Για τον υπολογισμό του p-value, συλλέγουμε το δείγμα και υπολογίζουμε τον κατάλληλο στατιστικό έλεγχο. Για παράδειγμα, t-statistic για έλεγχο των μέσων, Chi-Square ή F-statistic για έλεγχο των διακυμάνσεων κτλ. Χρησιμοποιώντας τη θεωρητική κατανομή των στατιστικών ελέγχων, βρέστε την περιοχή δεξιά (`κάτω`) της καμπύλης (για συνεχείς μεταβλητές) στην κατεύθυνση(εις) της εναλλαχτικής υπόθεσης (H_1) βάση ενός πίνακα.

Παράδειγμα

Ποιο είναι το p-value για τη μηδενική υπόθεση ότι τα ασφάλιστρα ΔΕΝ αυξάνονται με τις απαιτήσεις.

Βάση της μηδενικής υπόθεσης υπολογίζουμε το t-statistic:

$$t = \frac{b}{\text{SE}} = \frac{0.059}{0.0113} = 5.2\,.$$

Από τους πίνακες παρατηρούμε ότι για 5 βαθμούς ελευθερίας το t value είναι 5.2 και βρίσκεται πάνω

από $T_{2.5\%} = 4.77$. Έτσι,

$$\text{p-value} < 0.0025.$$

Η πιο πάνω τιμή παρέχει πολύ λίγη αξιοπιστία στην μηδενική υπόθεση H_0 και έτσι μπορούμε να την απορρίψουμε και να συμπεράνουμε ότι τα ασφάλιστρα αυξάνονται όσο αυξάνονται οι απαιτήσεις.

Σημειώστε ότι η εναλλακτική υπόθεση για το πιο πάνω παράδειγμα είναι ότι τα ασφλάλιστρα αυξάνονται με τις απαιτήσεις. Δηλαδή,

$$H_1 : \quad \beta > 0.$$

Θεωρήστε τη μηδενική υπόθεση ότι *τα ασφάλιστρα δεν συσχετίζονται με τις απαιτήσεις* (δηλαδή Y είναι ασυσχέτιστο με το X). Αυτό υπονοεί ότι η εναλλακτική υπόθεση λέει ότι τα ασφάλιστρα σχετίζονται με τις απαιτήσεις είτε με θετικό είτε με αρνητικό τρόπο. Έτσι, μπορούμε να γράψουμε αυτή την εναλλακτική υπόθεση ως:

$$H_1 : \quad \beta > 0 \text{ ή } \beta < 0,$$

ή αντίστοιχα

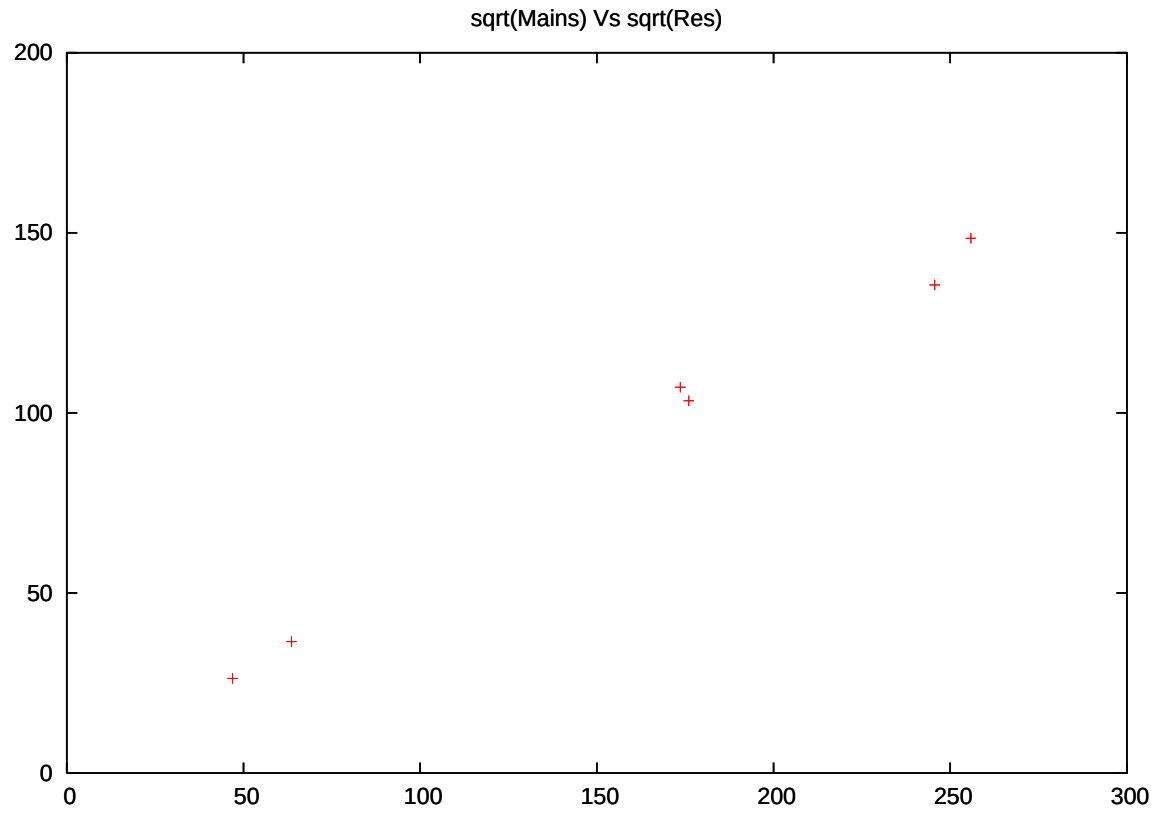
$$H_1 : \quad \beta \neq 0.$$

Αυτή είναι υπόθεση διπλής όψεως και πρέπει να υπολογιστεί p-value διπλής όψεως.

Ειδικές περιπτώσεις

Ο πιο κάτω πίνακας δείχνει τον πληθυσμό ζωνών (*Res.*) και τον αριθμό των οικιακών κεντρικών αγωγών (*Mains*). Επιθυμούμε να βρούμε πως το μέγεθος του πληθυσμού επηρεάζει τον αριθμό των τηλεφώνων. Διάφορα μοντέλα που συνδέουν αυτές τις δύο μεταβλητές έχουν χρησιμοποιηθεί για την εκτίμηση του πληθυσμού σε μικρές περιοχές σε χρονιές μη-απογραφής.

# Res.	4041	2200	30148	60324	65468	30988
# Mains	1332	690	11476	18368	22044	10686
# $\sqrt{\text{Res.}}$	63.57	46.90	173.63	245.60	255.87	176.03
# $\sqrt{\text{Mains}}$	36.50	26.27	107.13	135.53	148.47	103.37



Θέτουμε $y = \sqrt{\#\text{ των τηλεφώνων}}$ και $x = \sqrt{\text{μέγεθος του πληθυσμού}}$. Το γράφημα παρουσιάζει μια γραμμική σχέση με τη γραμμή να περνά από το σημείο $(0,0)$[α΄].

Θεωρήστε την πολινδρόμηση:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

όπου $\varepsilon_i \sim N(0, \sigma^2)$, και έτσι, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Οι εκτιμήσεις ελαχίστων τετραγώνων του β_0 και β_1

[α΄] Αυτό είναι ευνόητο λόγω του ότι όταν μια περιοχή δεν έχει πληθυσμό, τότε δεν θα υπάρχουν και γραμμές τηλεφώνου.

ορίζονται ως $b_0 = \bar{y} - b_1\bar{x}$ και $b_1 = \sum xy / \sum x^2$. Το b_i είναι γραμμικός συνδυασμός των y_i΄ς καθώς επίσης είναι και κανονικά κατανεμημένο.

Θέτουμε το SE (standard error) του b_j να ορίζεται ως $\mathrm{SE}(b_j)$. Τότε

$$\frac{(b_j - \beta_j)}{\mathrm{SE}(b_j)} \sim \mathrm{T}^{(n-2)}.$$

Επιπλέον, το $(1-a)\times 100$ ποσοστό Δ.Ε. για β_j δίνεται από

$$b_j - \mathrm{SE}(b_j)\,\mathrm{T}^{(n-2)}_{\frac{a}{2}} < \beta_j < b_j + \mathrm{SE}(b_j)\,\mathrm{T}^{(n-2)}_{\frac{a}{2}},$$

όπου $j = 0, 1$ και $\mathrm{T}^{(n-2)}_{\frac{a}{2}}$ είναι το άνω $a/2$ σημείο της t-κατανομής με $n-2$ βαθμούς ελευθερίας.

Computer output:

Διακύμανση	j	b_j	$\mathrm{SE}(b_j)$	$t(b_j)$	$P(\lvert t\rvert > \lvert t(b_j)\rvert)$
Τεταγμένη	0	1.301	4.280	0.3037	0.7763
$\sqrt{\text{Mains}}$	1	0.571	0.024	23.955	0.0001

Δηλαδή,

$b_0 = 1.301,\ b_1 = 0.571,\ \mathrm{SE}(b_0) = 4.28$ και $\mathrm{SE}(b_1) = 0.024.$

Επίσης,

$$t(b_0) = \frac{b_0}{\text{SE}(b_0)} = 0.3037 \quad \text{και} \quad t(b_1) = \frac{b_1}{\text{SE}(b_1)} = 23.955.$$

Καθώς το $T^{(4)}_{5\%} = 2.1318$, το 90% διάστημα εμπιστοσύνης για β_0 και β_1 είναι δοδομένα, ως αντιστοίχως:

$$(-7.8241, 10.4241) \qquad \text{και} \qquad (0.5198, 0.6221).$$

Το διάστημα του β_0 υποθέτει ότι β_1 είναι σταθερό και το αντίθετο.

Καθώς το 0 περιέχεται στο διάστημα του β_0 δεν μπορούμε να απορρίψουμε την $H_0 : \beta_0 = 0$. Όμως μπορούμε να απορρίψουμε την $H_0 : \beta_1 = 0.7$.

Η πιθανότητα η τιμή της t-κατανεμημένης τυχαίας μεταβλητής να είναι μαθηματικά μεγαλύτερη του $|t(b_0)| = 0.3037$ είναι 0.7763 και η πιθανότητα να έχουμε t-τιμή μεγαλύτερη του $|t(b_1)| = 23.995$ είναι 0.0001. Έτσι, μπορούμε να απορρίψουμε την $H_0 : \beta_1 = 0$ σε 5, 1, ή 0.1 τοις εκατό. Όμως δεν μπορούμε να απορρίψουμε την $H_0 : \beta_0 = 0$ για κάθε επίπεδο σημαντικότητας.

Όταν η τετμημένη (β_0) δεν δίνεται το αποτέλεσμα του υπολογιστή δίνεται από:

Διακύμανση	j	b_j	$\mathrm{SE}(b_j)$	$t(b_j)$	$P(\|t\| > \|t(b_j)\|)$
$\sqrt{\text{Mains}}$	1	0.578	0.0097	59.566	0.0001

Το $T^{(5)}_{5\%} = 2.0151$ και κατά συνέπεια το 90% Δ.Ε. για το β_1 δίνεται από:

$$(0.5583, 0.5973).$$

Προσαρμοστικότητα (Goodness of fit)

Ο συντελεστής προσδιορισμού R^2 (αναφέρεται ως R *Τετράγωνο*) μετρά την *ποιότητα προσαρμογής* της γραμμής παλινδρόμησης.

Θεωρήστε τις ακόλουθες ορολογίες:

- Total Sum of Squares (TSS): $\sum (Y - \bar{Y})^2$.
- Regression Sum of Squares (RSS): $\sum (\widehat{Y} - \bar{Y})^2$.
- Sum of squares of errors (SSE): $\sum (Y - \widehat{Y})^2$.

Έχουμε:

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{SSE},$$

$$\boxed{R^2 = \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{RSS}}{\mathrm{RSS} + \mathrm{SSE}} = 1 - \frac{\mathrm{SSE}}{\mathrm{TSS}}}$$

και

$$\boxed{0 \leq R^2 \leq 1.}$$

$R^2 = 1$ υποδηλώνει ότι οι δειγματικές παρατηρήσεις βρίσκονται ακριβώς πάνω στη γραμμή παλινδρόμησης, καθώς $R^2 = 0$ υποδηλώνει ότι η γραμμή παλινδρόμησης δεν έχει καμία χρήση. Δηλαδή το X δεν επηρεάζει καθόλου το Y (γραμμικά).

Παράδειγμα

Χρησιμοποιώντας το παράδειγμα *Απαιτήσεις-Ασφάλιστρα* , TSS $= 1150$, SSE $= 177.68$, και έτσι

$$R^2 = \frac{1150 - 177.68}{1150} = 0.845.$$

Η ερμηνεία είναι: 84.5% της διακύμανσης στα ασφάλιστρα (Y) εξηγείται από τη διακύμανση στις απαιτήσεις (X). Το ποσοτό αυτό είναι αξιοσέβαστο, μιας και αφήνει μόλις 15.5% της διακύμανσης να εξηγείται από άλλους παράγοντες.

Σημειώστε (θα μπορούσε να ειπωθεί νωρίτερα):

Ο εκτιμητής ελαχίστων τετραγώνων του β_0 στην απλή παλινδρόμηση $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ δίνεται από $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Έτσι η διακύμανση του $\hat{\beta}_0$ δίνεται από:

$$\begin{aligned}\mathrm{Var}(\beta_0) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) = \mathrm{Var}(\bar{y}) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) \\ &= \frac{1}{n^2} \sum \mathrm{Var}(y_i) + \bar{x}^2 \frac{\sigma^2}{\sum x_i^2} = \frac{n\sigma^2}{n^2} + \bar{x}^2 \frac{\sigma^2}{\sum x_i^2} \\ &= \sigma^2 \Big(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2} \Big).\end{aligned}$$

Έτσι, $\mathrm{SE}(\hat{\beta}_0) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2}}$.

Έλεγχος συνολικής σημαντικότητας

Θεωρήστε την παλινδρόμηση:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Για τον έλεγχο της γραμμικής συσχέτισης:

$$\begin{aligned} \mathrm{H}_0 : &\quad \beta_0 = \beta_1 = 0 \\ \mathrm{H}_1 : &\quad \beta_0 \neq 0 \quad \text{και} \quad \beta_1 \neq 0 \end{aligned}$$

Για τον έλεγχο υπολογίζουμε το F-στατιστικό:

$$F = \frac{\mathrm{RSS}}{\mathrm{SSE}/(n-2)}$$

Απορρίπτουμε την H_0 όταν F υπερβαίνει $F_{\alpha\%}^{(1,n-2)}$, όπου $(1, n-2)$ είναι οι βαθμοί ελευθερίας της F κατανομής και α επιλεγμένο (επί τοις εκατό) σημείο.

Στο παράδειγμα *απαιτήσεις και ασφάλιστρα* το F-στατιστικό υπολογίζεται από

$$\frac{\mathrm{RSS}}{\mathrm{SSE}/(n-2)} = \frac{972.32}{177.67/5} = 27.36.$$

Το $F_{2.5\%}^{(1,5)} = 10.01$. Έτσι, το H_0 απορρίπτεται.

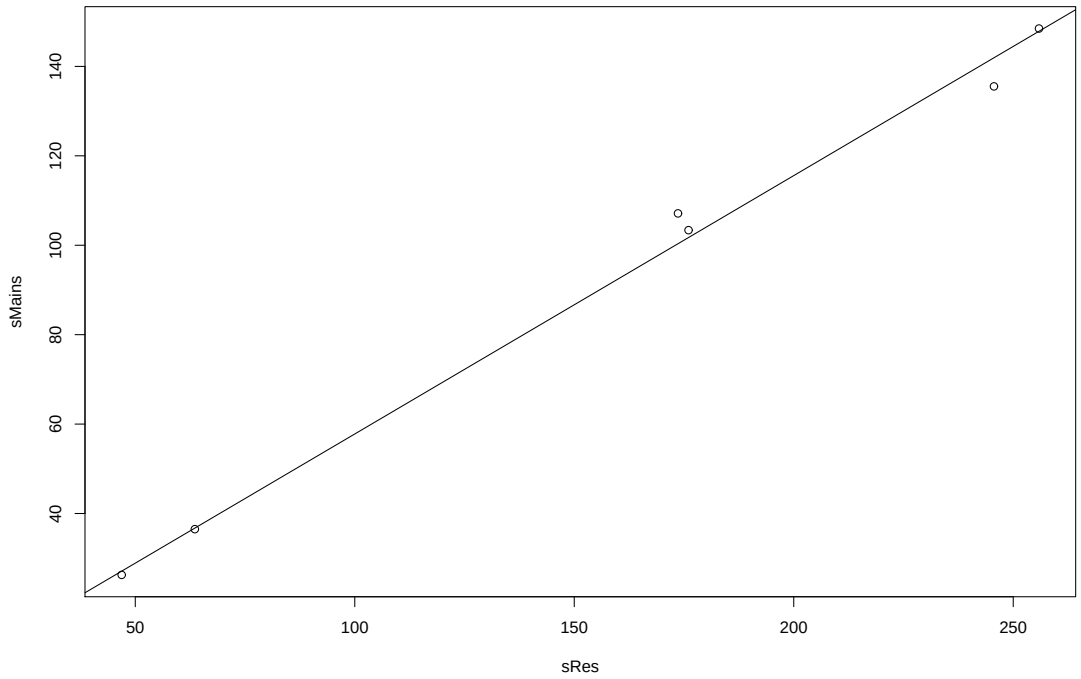
Εκτίμηση υπολογιστή (Mains and Residence)

```
Res=c(4041, 2200, 30148, 60324, 65468, 30988)
Mains=c(1332, 690, 11476, 18368, 22044, 10686)
sRes=sqrt(Res)      sMains=sqrt(Mains)

> reg <- lm(sMains~sRes)
sRes =   63.57  46.90 173.63 245.61 255.87 176.03
sMains = 36.50  26.27 107.13 135.53 148.47 103.37
Residuals: 1      2     3     4     5     6
         -1.13 -1.83  6.61 -6.11  0.97  1.49
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30    4.28   0.30    0.776
sRes         0.57    0.02  23.96  1.8e-05 ***
---
Residual standard error: 4.71 on 4 DF
Multiple R-Squared: 0.99, Adjusted R-squared: 0.99
F-statistic: 573.8 on 1 and 4 DF,  p-value: 1.8e-05

> reg <- lm(sMains~sRes-1)
Residuals: 1      2     3    4     5      6
         -0.24 -0.84  6.79 -6.40  0.61  1.65
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
sRes 0.58   0.010   59.56 2.53e-08 ***
---
Residual standard error: 4.26 on 5 DF
```

```
Multiple R-Squared: 1.0, Adjusted R-squared: 0.99
F-statistic:  3547 on 1 and 5 DF,  p-value: 2.5e-08
```



Εκτίμηση υπολογιστή (Απαιτήσεις και Ασφάλιστρα)

```
> x=c(100, 200, 300, 400,500, 600, 700);
> y=c(40, 50, 50, 70,65, 65, 80);
Residuals:
  1     2     3     4     5     6     7
-2.32  1.79 -4.11 10.00 -0.89 -6.79  2.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.038   7.231  0.00079
```

```
x            0.059    0.011   5.231  0.00338
---
Residual standard error: 5.961 on 5 df
Multiple R-Squared: 0.85, Adjusted R-squared: 0.81
F-statistic: 27.36 on 1 and 5 DF, p-value: 0.003
```

Πολλαπλή Παλινδρόμηση

Source: Long-Kogan Realty, Chicago, USA.

y	PRICE	Selling price of house in thousands of dollars
X_1	BDR	Number of bedrooms
X_2	FLR	Floor space in sq.ft.
X_3	FP	Number of fireplaces
X_4	RMS	Number of rooms
X_5	ST	Storm windows (1 if present, 0 if absent)
X_6	LOT	Front footage of lot in feet
X_7	TAX	Annual taxes
X_8	BTH	Number of bathrooms
X_9	CON	Construction (0 if frame, 1 if brick)
X_{10}	GAR	Garage size (0 = no garage, 1 = one-car garage, etc.)
X_{11}	CDN	Condition (1 = 'need work', 0 otherwise)
X_{12}	L1	Location (L1 = 1 if property is in zone A, L1 = 0 otherwise)
X_{13}	L2	Location (L2 = 1 if property is in zone B, L2 = 0 otherwise)

Price = f(FLR, ST, LOT, CON, GAR, L2)

26 x 13 (26 observations 13 independent variables)

Y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
53	2	967	0	5	0	39	652	1.5	1	0.0	0	1	0
55	2	815	1	5	0	33	1000	1.0	1	2.0	1	1	0
56	3	900	0	5	1	35	897	1.5	1	1.0	0	1	0
58	3	1007	0	6	1	24	964	1.5	0	2.0	0	1	0
64	3	1100	1	7	0	50	1099	1.5	1	1.5	0	1	0
44	4	897	0	7	0	25	960	2.0	0	1.0	0	1	0
49	5	1400	0	8	0	30	678	1.0	0	1.0	1	1	0
70	3	2261	0	6	0	29	2700	1.0	0	2.0	0	1	0
72	4	1290	0	8	1	33	800	1.5	1	1.5	0	1	0
82	4	2104	0	9	0	40	1038	2.5	1	1.0	1	1	0
85	8	2240	1	12	1	50	1200	3.0	0	2.0	0	1	0
45	2	641	0	5	0	25	860	1.0	0	0.0	0	0	1
47	3	862	0	6	0	25	600	1.0	1	0.0	0	0	1
49	4	1043	0	7	0	30	676	1.5	0	0.0	0	0	1
56	4	1325	0	8	0	50	1287	1.5	0	0.0	0	0	1
60	2	782	0	5	1	25	834	1.0	0	0.0	0	0	1
62	3	1126	0	7	1	30	734	2.0	1	0.0	1	0	1
64	4	1226	0	8	0	37	551	2.0	0	2.0	0	0	1
66	2	929	0	5	0	30	1355	1.0	1	1.0	0	0	1
35	4	1137	1	7	0	25	561	1.5	0	0.0	0	0	0
38	3	743	0	6	0	25	489	1.0	1	0.0	0	0	0
43	3	596	0	5	0	50	752	1.0	0	0.0	0	0	0
46	2	803	0	5	0	27	774	1.0	1	0.0	1	0	0
46	2	696	0	4	0	30	440	2.0	1	1.0	0	0	0
50	2	691	0	6	0	30	549	1.0	0	2.0	1	0	0
65	3	1023	0	7	1	30	900	2.0	1	1.0	0	1	0

Μοντέλο παλινδρόμησης σε μορφή πινάκων

Υποθέστε ότι υπάρχουν n ανεξάρτητες μεταβλητές και m παρατηρήσεις. Το μοντέλο παλινδρόμησης εκφράζεται ως:

$$\begin{aligned}
y_1 =&\beta_0+ \beta_1 x_{11}+ \beta_2 x_{12}+\cdots+ \beta_n x_{1n}+ \varepsilon_1\\
y_2 =&\beta_0+ \beta_1 x_{21}+ \beta_2 x_{22}+\cdots+ \beta_n x_{2n}+ \varepsilon_2\\
\vdots \quad &\vdots \qquad\qquad\qquad\qquad\qquad \vdots\\
y_m =&\beta_0+\beta_1 x_{m1}+\beta_2 x_{m2}+\cdots+\beta_n x_{mn}+\varepsilon_m
\end{aligned}$$

Η i παρατήρηση εκφράζεται ως:

$$y_i = \begin{pmatrix}1 & x_{i1} & x_{i2} & \cdots & x_{in}\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_n\end{pmatrix}+\varepsilon_i$$

και το όλο σύστημα των παρατηρήσεων εκφράζεται ως:

$$\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{pmatrix}=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1n}\\ 1 & x_{21} & x_{22} & \cdots & x_{2n}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn}\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_n\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_m\end{pmatrix},$$

or

$$y = X\beta + \varepsilon.$$

Παράδειγμα (Απαιτήσεις και Ασφάλιστρα)

Θεωρήστε το απλό μοντέλο παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

όπου X δειλώνει τις απαιτήσεις, Y τα ασφάλιστρα και

x	100	200	300	400	500	600	700
y	40	50	50	70	65	65	80

Η παλινδρόμηση σε μορφή πινάκων δίνεται ως εξής:

$$\begin{pmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{pmatrix} = \begin{pmatrix} 1 & 100 \\ 1 & 200 \\ 1 & 300 \\ 1 & 400 \\ 1 & 500 \\ 1 & 600 \\ 1 & 700 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{pmatrix}$$

Το πιο πάνω είναι ισότιμο με

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_7)$$

όπου

$$y = \begin{pmatrix} 40 \\ 50 \\ 50 \\ 70 \\ 65 \\ 65 \\ 80 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 100 \\ 1 & 200 \\ 1 & 300 \\ 1 & 400 \\ 1 & 500 \\ 1 & 600 \\ 1 & 700 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{και} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{pmatrix}.$$

Εκτιμητές μεθόδου ελαχίστων τετραγώνων

- Θεωρήστε το πολλαπλό γραμμικό μοντέλο παλινδρόμησης:

$$y = X\beta + \varepsilon, \tag{1}$$

όπου $y \in \Re^m$, $X \in \Re^{m\times(n+1)}$, $\beta \in \Re^{(n+1)}$ και $\varepsilon \in \Re^m$.

- Η πιο διαδεδομένη τεχνική εκτίμησης του μοντέλου (2) είναι τα ελάχιστα τετράγωνα.

- Θέτουμε b^* ένα αυθαίρετο διάνυσμα n-στοιχείων και το αντίστοιχο διάνυσμα σφαλμάτων $e^* \in \Re^m$ ορίζεται

από:

$$e^* = y - Xb^* \,. \tag{2}$$

- Ο *κανόνας ελαχίστων τετραγώνων* για την επιλογή του $b^* = (b_0^* \dots b_n^*)^T$ προϋποθέτει την ελαχιστοποίηση του *αθροίσματος των τετραγώνων των σφαλμάτων*, ή *Residuals sum of squared* (RSS):

$$\begin{aligned} \|e\|_2^2 &= e^{*T} e^* \\ &= \sum_{i=1}^{m} e_i^2 = e_1^2 + \cdots e_m^2 \end{aligned}$$

- Έτσι, η εκτίμηση ελαχίστων τετραγώνων προέρχεται από την ελαχιστοποίηση:

$$\begin{aligned} e^{*T} e^* &= \sum_{i=1}^{n} (y_i - b_0^* - b_1^* x_{11} - b_2^* x_{12} - \cdots - b_n^* x_{1n})^2 \\ &= (y - Xb^*)^T (y - Xb^*) \\ &= y^T y - b^{*T}(X^T y) - y^T X b^* + b^{*T}(X^T X) b^* \\ &= y^T y - 2b^{*T}(X^T y) + b^{*T}(X^T X) b^* \end{aligned}$$

- Σημειώστε ότι όλες οι ποσότητες είναι αριθμητικά

μεγέθη, οπόταν

$$b^{*T}(X^T y) = y^T X b^*.$$

Κατά συνέπεια, η εκτίμηση ελαχίστων τετραγώνων του β, ας είναι $\widehat{\beta}$, προέρχεται από τη λύση:

$$\begin{aligned}\widehat{\beta} &= \underset{b^*}{\operatorname{argmin}} \left\|e\right\|_2^2 = \underset{b^*}{\operatorname{argmin}}\ e^{*T}e^* \\ &= \underset{b^*}{\operatorname{argmin}} \Big(y^T y - 2b^{*T}(X^T y) + b^{*T}(X^T X)b^*\Big).\end{aligned}$$

- Για την ελαχιστοποίηση $e^{*T}e^*$ παραγοντοποιούμε ως προς το b^* και το θέτουμε ίσο με μηδέν.

- Έτσι,

$$\frac{d(e^{*T}e^*)}{db^*} = -2(X^T y) + 2(X^T X)b^*.$$

Θέτοντας το πιο πάνω ίσο με 0 παίρνουμε τις OLS *κανονικές εξισώσεις*:

- Εάν υποθέσουμε ότι οι ανεξάρτητες μεταβλητές είναι γραμμικά ασυσχέτιστες, δηλαδή, ο X έχει πλήρη τάξη, τότε $(X^T X)$ έχει αντίστροφο.

- Πολλαπλασιάζοντας κάθε πλευρά με $(X^T X)^{-1}$ δίνει

$$(X^T X)^{-1}(X^T X)\widehat{\beta} = (X^T X)^{-1} X^T y,$$

ή η αντίστοιχη πρόταση:

$$\boxed{\widehat{\beta} = (X^T X)^{-1} X^T y.}$$

- Ο OLS εκτιμητής $\widehat{\beta}$ είναι μοναδικός.

Παραδείγματα

Στο παράδειγμα των *Απαιτήσεων και των Ασφαλίστρων* το διάνυσμα y και ο πίνακας X δίνονται από:

$$y^T = (40\ \ 50\ \ 50\ \ 70\ \ 65\ \ 65\ \ 80) \quad \text{και}$$

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 100 & 200 & 300 & 400 & 500 & 600 & 700 \end{pmatrix}.$$

Έτσι,

$$X^T X = \begin{pmatrix} 7 & 2800 \\ 2800 & 1400000 \end{pmatrix}, \qquad X^T y = \begin{pmatrix} 420 \\ 184500 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{196} \begin{pmatrix} 140 & -2.8 \\ -2.8 & 7 \times 10^{-4} \end{pmatrix} \quad \text{και}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} 36.43 \\ 0.059 \end{pmatrix}.$$

Γενικά, αν $n = 2$, τότε:

$$X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

Σημειώστε ότι ο αριθμός κατάστασης του X δίνεται από $\text{Cond}(X) = 1000.0$. Αν η μεταβλητή x (απαιτήσεων) διαιρεθεί με 100, τότε ο αριθμός κατάσταση γίνεται 10.404.

Παράδειγμα

Υπάρχουν διάφορα στατιστικά πακέτα για τον υπολογισμό των ελαχίστων-τετραγώνων και άλλων ποσοτήτων. Το πακέτα αυτά είναι SPSS, SAS, GLIM, S-PLUS, R, EXCEL, κτλ. Για το παράδειγμα *House prices* η εξίσωση παλινδρόμησης (χωρίς τη

χρήση όλων των μεταβλητών) δίνεται από:

$$\text{PRICE} = 18.48 + 0.18\,\text{FLR} + 4.03\,\text{RMS} - 7.75\,\text{BDR} + 2.20\,\text{BTH} + 1.37\,\text{GAR} + 0.257\,\text{LOT} + 7.09\,\text{FP} + 10.96\,\text{ST}.$$

Θεωρήστε την εκτιμημένη τιμή πώλησης ενός σπιτιού με 1000 τετραγωνικά πόδια έκταση, 8 δωμάτια, 4 υποπνοδωμάτια, 2 μπάνια, ένα μονωτικό παράθυρο κατά καταιγιδών, χωρίς τζάκια, 40 πόδια μπροστινή αυλή και 1 γκαράζ:

$$18.48 + 0.18(1000) + 4.03(8) - 7.75(4) + 2.20(2) + 1.37(1) + 0.257(40) + 7.09(0) + 10.96(1) = 64.73.$$

Από την παλινδρόμηση μπορεί να διαπιστωθούν τα ακόλουθα:

- Ένα επιπλέον αυτοκίνητο στο γκαράζ αυξάνει την τιμή του σπιτιού κατά \$1370.
- Αύξηση του εσωτερικού χώρου κατά ένα τετραγωνικό πόδι αυξάνει την τιμή κατά \$18.

Κάθως η μεταβολή στην τιμή είναι οριακή, δηλαδή, οι άλλες παραμέτροι μένουν σταθερές.

Παρατηρήστε το αρνητικό πρόσημο που σχετίζεται με τον αριθμό των δωματίων (BDM). Αυτό υπονοεί

ότι η αξία του σπιτιού μειώνεται όταν αυξάνεται ο αριθμός των υπνοδωματίων χωρίς να αυξάνονται ταυτόχρονα η επιφάνεια και ο αριθμός των συνολικών δωματίων. Π.χ. αν προστεθεί ένα δωμάτιο, ένα υπνοδωμάτιο και επιφάνεια σπιτιού, τότε η εκτιμώμενη τιμή αυξάνεται.

Στις περιπτώσεις όπου υπάρχουν διάφορες σχετιζόμενες μεταβλητές, τα πρόσημά τους είναι κοινά. Περαιτέρω έρευνα επεξηγεί το φαινόμενο αυτό.

Επιπλέον, οι εκτιμήσεις είναι τυχαίες μεταβλητές ενώ πιθανόν να μην έχουμε επιλέξει τις πιο σημαντικές επεξηγηματικές μεταβλητές. Άρα το μοντέλο μας μπορεί να μην δίνει την πραγματική `σωστή` εικόνα τις κατάστασεις.

Η δημιουργία τέλειου μοντέλου είναι σχεδόν απίθανη.

Υποθέσεις τυπικού μοντέλου γραμμ. παλινδρ.

Θεωρήστε την παλινδρόμηση:

$$y_i = X_i\beta + \varepsilon_i \quad \text{or} \quad y = X\beta + \varepsilon.$$

Για να ισχύουν οι στατιστικές ιδιότητες των εκτιμητών του β χρειάζεται να γίνουν κάποιες υποθέσεις σχετικά με το πως δημιουργούνται οι παρατηρήσεις y.

- $\mathrm{E}(\varepsilon) = 0$. Δηλαδή, $\mathrm{E}(y) = X\beta$.

Υποθέστε ότι X μεταβλητές μετρούν το οικογενειακό εισόδημα και άλλα οικογενειακά χαρακτηριστικά και Y υποδηλώνει τα ταξιδιωτικά έξοδα της οικογένειας. Η πρώτη γραμμή του X πίνακα δίνει τα ποσά εισοδήματος, μεγέθους και σύνθεσης της οικογένειας. Θέτουμε s_1 το διάνυσμα γραμμής που αποτελείται από τρεις αριθμούς. Το μέσο ή αναμενόμενο επίπεδο ταξιδιωτικών εξόδων για αυτή την οικογένεια δίνεται από:

$$\mathrm{E}(y_1) = s_1\beta.$$

Εντούτοις, τα *πραγματικά* ταξιδιωτικά έξοδα οικογενειών με αυτά τα χαρακτηριστικά, μπορέι να είναι μεγαλύτερα ή μικρότερα των αναμενόμενων. Επιπλέον, τα έξοδα τις ίδιας οικογένειας κυμαίνονται γύρω από τη μέση τιμή σε διαφορετικές περιόδους. *Αν όλες οι σημαντικές μεταβλητές περιλαμβάνονται στο X, τότε αναμένουμε ότι οι μέσες θετικές και αρνητικές διακυμάνσεις από την αναμενόμενη τιμή θα είναι μηδέν.* Δηλαδή, $\mathrm{E}(\varepsilon_1) = 0$.

Παρόμοιες εκτιμήσεις ισχύουν για κάθε γραμμή του X, και έτσι έχουμε:

$$\mathrm{E}(\varepsilon) = \begin{pmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_m) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

- $\mathrm{E}(\varepsilon\varepsilon^T) = \sigma^2 I_m$.

  Ο πίνακς διακυμάνσεων του ε δίνεται από $\mathrm{E}\Big((\varepsilon - \mathrm{E}(\varepsilon))((\varepsilon - \mathrm{E}(\varepsilon))^T\Big) = \mathrm{E}(\varepsilon\varepsilon^T)$ καθώς $\mathrm{E}(\varepsilon) = 0$.

Δηλαδή,

$$\mathrm{E}(\varepsilon\varepsilon^T) = \begin{pmatrix} \mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1,\varepsilon_2) & \dots & \mathrm{Cov}(\varepsilon_1,\varepsilon_m) \\ \mathrm{Cov}(\varepsilon_2,\varepsilon_1) & \mathrm{Var}(\varepsilon_2) & \dots & \mathrm{Cov}(\varepsilon_2,\varepsilon_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\varepsilon_m,\varepsilon_1) & \mathrm{Cov}(\varepsilon_m,\varepsilon_2) & \dots & \mathrm{Var}(\varepsilon_m) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 I_m$$

Το πιο πάνω είναι διπλή υπόθεση, δηλαδή:

1. Κάθε ε_i κατανομή έχει την ίδια διακύμανση.

   Αυτή η ιδιότητα αναφέρεται ως *ομοσκεδαστικότητα (homoscedasticity)* (ομοιογενείς διακυμάνσεις). Το αντίθετο φαινόμενο ονομάζεται *ετεροσκεδαστικότητα (heteroscedasticity)*.

   Π.χ. Αν θεωρήσουμε μια διατομή πληθυσμού, τότε η υπόθεση *ετεροσκεδαστικότητας* είναι πιο λογική. Αυτό γιατί οικογένειες με χαμηλό εισόδημα θα έχουν με σιγουριά λιγότερα ταξιδιωτικά έξοδα κατά μέσο όρο. Επιπλέον θα υπάρχει μικρότερη διακύμανση αυτών των εξόδων γύρω από το μέσο. Από την άλλη

οικογένειες με ψηλά εισοδήματα τείνουν να παρουσιάζουν τόσο ψηλότερα επίπεδα μέσων εξόδων όσο και μεγαλύτερη διακύμανση γύρω από το μέσο.

2. Όλες οι διακυμάνσεις είναι κατά ζεύγη ασυσχέτιστες.

   Αυτή είναι μια ισχυρή υπόθεση. Υπονοεί για παράδειγμα ότι ψηλά έξοδα σε μια χρονιά δεν σχετίζονται με αντίστοιχα χαμηλά (ή ψηλά) έξοδα την επόμενη χρονιά, ή τις επόμενες χρονιές. Ακόμη ένα παράδειγμα είναι ότι η πιο πάνω υπόθεση *δεν λαμβάνει υπόψη τον παράγοντα της γειτονιάς*. Δηλαδή, το μέγεθος της διακύμανσης των εξόδων για μια οικογένεια δεν επιρεάζει το μέγεθος της διακύμανσης για άλλη οικογένεια.

- Το X είναι ένας μη-στοχαστικός πίνακας: $\mathrm{E}(X^T\varepsilon) = 0$.

  Αυτό σημαίνει ότι αν χρησιμοποιήσουμε ένα άλλο δείγμα n παρατηρήσεων, τότε ο X πίνακας των επεξηγηματικών μεταβλητών μένει ανέπαφος. Η μόνη πηγή διακύμανσης είναι στο ε και κατά συνέπεια στο y.

Μέσος και διακύμανση εκτιμητών

Θεωρήστε την παλινδρόμηση:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

Ο εκτιμητής OLS δίνεται από:

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

Αντικαθιστώντας $y = X\beta + \varepsilon$ στο τελευταίο δίνει:

$$\begin{aligned}
\widehat{\beta} &= (X^T X)^{-1} X^T y \\
&= (X^T X)^{-1} X^T (X\beta + \varepsilon) \\
&= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \varepsilon \\
&= \beta + (X^T X)^{-1} X^T \varepsilon.
\end{aligned}$$

Έτσι,

$$\begin{aligned}\mathrm{E}(\widehat{\beta}) &= \mathrm{E}(\beta + (X^TX)^{-1}X^T\varepsilon)\\ &= \mathrm{E}(\beta) + (X^TX)^{-1}X^T\mathrm{E}(\varepsilon)\\ &= \beta.\end{aligned}$$

Σημειώστε ότι
$\widehat{\beta} - \mathrm{E}(\widehat{\beta}) = \widehat{\beta} - \beta = (X^TX)^{-1}X^T\varepsilon.$

Έτσι,

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}) &= \mathrm{E}\Big((\widehat{\beta} - \mathrm{E}(\widehat{\beta}))(\widehat{\beta} - \mathrm{E}(\widehat{\beta}))^T\Big)\\ &= \mathrm{E}\Big((X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}\Big)\\ &= (X^TX)^{-1}X^T\mathrm{E}(\varepsilon\varepsilon^T)X(X^TX)^{-1}\\ &= (X^TX)^{-1}X^T\sigma^2I_mX(X^TX)^{-1}\\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}\\ &= \sigma^2(X^TX)^{-1}.\end{aligned}$$

Τα στοιχεία της βασικής διαγωνίου του
$\mathrm{Var}(\widehat{\beta}) = \sigma^2(X^TX)^{-1}$ καθορίζουν τις αντίστοιχες διακυμάνσεις των στοιχείων του $\widehat{\beta}$.

Εκτίμηση σ^2

Συνήθως σ^2 είναι άγνωστο και χρειάζεται να

εκτιμηθεί για τη διεξαγωγή συμπερασμάτων. Αυτό μπορεί να γίνει χρησιμοποιώντας τα σφάλματα (υπόλοιπα) e_i. Ένας αμερόληπτος εκτιμητής του σ^2 δίνεται από:

$$\boxed{s^2 = \frac{1}{m-n-1}\sum_{i=1}^{n} e_i^2 = \frac{e^T e}{m-n-1}.}$$

Παράδειγμα (Απαιτήσεις και Ασφάλιστρα)

Ο εκτιμητής του σ^2 είναι $S^2 = 35.53$. Επιπλέον,

$$X^T X = \begin{pmatrix} 7 & 2800 \\ 2800 & 1400000 \end{pmatrix}$$

και

$$S^2(X^T X)^{-1} = 35.53 \times \frac{1}{196}\begin{pmatrix} 140 & -2.8 \\ -2.8 & 7\times 10^{-4} \end{pmatrix}$$

$$= \begin{pmatrix} 25.38 & -0.051 \\ -0.051 & 0.0001 \end{pmatrix}.$$

Τα διαγώνια στοιχεία του $S\sqrt{(X^T X)^{-1}}$ δίνονται από:

$$5.038 \quad \text{και} \quad 0.011.$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.038   7.231  0.00079
x            0.059    0.011   5.231  0.00338
---
Residual standard error: 5.961 on 5 df
```

Παράδειγμα (House data)

Θεωρήστε τον υπολογισμό του μοντέλου:

$$\text{PRICE} = \beta_1 \text{FLR} + \beta_2 \text{RMS} + \beta_3 \text{BDR} + \beta_4 \text{GAR} + \beta_5 \text{ST} + \varepsilon.$$

Η μεταβλητή απόκρισης $y = \text{PRICE}$ και ο 26×5 εξωγενής πίνακας X δίνονται από:

$$y = \begin{pmatrix} 53 \\ 55 \\ \vdots \\ 65 \end{pmatrix} \quad \text{και} \quad X = (\text{FLR RMS BDR GAR ST}) = \begin{pmatrix} 967 & 5 & 2 & 0.0 & 0 \\ 815 & 5 & 2 & 2.0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1023 & 7 & 3 & 1.0 & 1 \end{pmatrix}$$

Τώρα,

$$X^T y = \begin{pmatrix} 1712260 \\ 9801 \\ 4884 \\ 1376 \\ 458 \end{pmatrix},$$

$$X^T X = \begin{pmatrix} 36714794 & 200359.0 & 102510.0 & 28014.0 & 8368.0 \\ 200359 & 1171.0 & 597.0 & 153.5 & 50.0 \\ 102510 & 597.0 & 314.0 & 77.5 & 26.0 \\ 28014 & 153.5 & 77.5 & 35.5 & 7.5 \\ 8368 & 50.0 & 26.0 & 7.5 & 7.0 \end{pmatrix}$$

και

$$(X^T X)^{-1} = \frac{1}{1000} \begin{pmatrix} 0.00 & -0.06 & -0.02 & -0.05 & 0.03 \\ -0.06 & 37.73 & -50.66 & -3.84 & -5.05 \\ -0.02 & -50.66 & 106.04 & 8.69 & -13.67 \\ -0.05 & -3.84 & 8.69 & 72.76 & -17.17 \\ 0.03 & -5.05 & -13.67 & -17.17 & 211.49 \end{pmatrix}$$

Έτσι, $\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{pmatrix} \equiv \hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} 0.015 \\ 11.401 \\ -12.519 \\ 3.040 \\ 9.413 \end{pmatrix}.$

Το υπόλοιπο $e = y - X\hat{\beta}$ και ο εκτιμητής του σ^2 υπολογίζονται από $S^2 = e^T e/(m-n)$, όπου $m = 26$ και $n = 5$. Δηλαδή,

$$S^2 = 62.98, \quad \text{ή} \quad S = 7.94.$$

Τα διαγώνια στοιχεία του $S\sqrt{(X^T X)^{-1}}$ δίνουν τα τυπικά σφάλαμτα του $\hat{\beta}$, δηλαδή.

$$\begin{pmatrix} 0.005 & 1.542 & 2.584 & 2.141 & 3.650 \end{pmatrix}^T.$$

Ο *ηλεκτρονικός υπολογισμός* δίνει:

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
FLR   0.015    0.005    2.78    0.011
RMS  11.401    1.542    7.40    2.8e-07
BDR -12.519    2.584   -4.85    8.7e-05
GAR   3.040    2.141    1.42    0.17
ST    9.413    3.650    2.58    0.02
---
Residual standard error: 7.94 on 21 DF
Multiple R-Squared: 0.98, Adjusted R-squared: 0.98
```

Θεώρημα Gauss-Markov

Ο εκτιμητής OLS $\widehat{\beta} = (X^T X)^{-1} X^T y$ είναι ο *Καλύτερος Γραμμικός Αμερόληπτος Εκτιμητής*, *Best Linear unbiased estimator* (BLUE). Αυτό υπονοεί ότι:

1. $\mathrm{E}(\widehat{\beta}) = \beta$.

   Η γραμμικότητα αναφέρεται στο y (ή ε). Δηλαδή, κάθε στοιχείο του $\widehat{\beta}$ είναι ένας γραμμικός συνδιασμός του y (ή ε).

2. Κανένας άλλος γραμμικός εκτιμητής δεν έχει μικρότερη δειγματική διακύμανση όπως αυτή των OLS εκτιμητών $\widehat{\beta}$.

Το θεώρημα Gauss-Markov δηλώνει ότι ο εκτιμητής ελαχίστων τετραγώνων του $\hat{\beta}$ είναι καλή επιλογή. Εντούτοις, αν τα σφάλματα συσχετίζονται ή έχουν άνιση διακύμανση, υπάρχουν καλύτεροι εκτιμητές. Σε κάποιες περιπτώσεις μη-γραμμικοί ή μεροληπτικοί εκτιμητές δουλεύουν καλύτερα. Έτσι, το θεώρημα δεν υπονοεί ότι κάποιος πρέπει πάντοτε να χρησιμοποιεί τα ελάχιστα τετράγωνα. Απλά προτείνονται ανεπιφύλακτα εκτός και αν υπάρχει

κάποιος σοβαρός λόγος για το αντίθετο. Π.χ.

1. Αν τα σφάλματα συσχετίζονται ή έχουν άνισες διακυμάνσεις, τότε τα γενικευμένα ελάχιστα τετράγωνα (generalized least-squares) μπορούν να χρησιμοποιηθούν.

2. Όταν οι εκτιμητές είναι άκρως συσχετισμένοι (συνευθειακοί, συγγραμμικοί), τότε μεροληπτικοί εκτιμητές όπως παλινδρόμηση κουφογραμμής (ridge regression) προτιμάται.

Ποιότητα προσαρμογής

Ένα στατιστικό μέτρο που χρησιμοποιείται ευρέως στον καθορισμό της προσαρμοστικότητας του μοντέλου παλινδρόμησης είναι ο συντελεστής προσδιορισμού R^2. Ο R^2 εξηγεί το ποσοστό της μεταβλητότητας στο y που μπορεί να εξηγηθεί από τη συσχέτιση με το X, δηλαδή, πόσο κοντά είναι οι παρατηρήσεις στην γραμμή παλινδρόμησης. Ο συντελεστή προσδιορισμού (Coefficient of Determination) $0 \leq R^2 \leq 1$ υπολογίζεται από όλα τα

στατιστικά πακέτα. Ορίζεται ως:

$$\boxed{R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}.}$$

Συχνά μικρά δείγματα διογκώνουν το R^2. Ο R^2 πάντοτε αυξάνεται με την προσθήκη μεταβλητών. Συγκεκριμένα, η προσθήκη μιας μεταβλητής στο μοντέλο μειώνει το RSS και αυξάνει το R^2.

Έτσι, R^2 από μόνο του δεν αποτελεί ικανοποιητικό κριτήριο αφού επιρεάζεται από το μέγεθος του μοντέλου. Όσο μεγαλύτερο το μοντέλο τόσο μεγαλύτερος ο συντελεστής.

Παράδειγμα (Απαιτήσεις και Ασφάλιστρα)

Οι απαιτήσεις και τα ασφάλιστρα δίνονται, αντίστοιχα, από:

(απαιτήσεις) x	100	200	300	400	500	600	700
(ασφάλιστρα) y	40	50	50	70	65	65	80

Ο *ηλεκτρονικός υπολογισμός* της γραμμικής παλινδρόμησης

$$y = \beta_0 + \beta_1 x + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

δίνει:

```
Residuals:
 1       2    3     4     5     6      7
-2.32  1.79 -4.11 10.00 -0.90 -6.79  2.32
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.428    5.04   7.23  0.00079 ***
x            0.059    0.01   5.23  0.00338 **
---
Residual standard error: 5.96 on 5 DF
Multiple R-Squared: 0.85, Adj R-squared: 0.81, cp=2
F-statistic: 27.36 on 1 and 5 DF,  p-value: 0.0033
```

Θεωρήστε την δημιουργία μιας τυχαίας μεταβλητής z από ομοιόμορφη κατανομή μεταξύ $\min = 100$ και

$\max = 1000$. Δηλαδή
$z^T =$
$(129.29, 231.47, 770.31, 127.14, 674.62, 217.54, 278.22)$.

Ο *ηλεκτρονικός υπολογισμός* της γραμμικής παλινδρόμησης

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

δίνει:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.765   5.97    6.46   0.003 **
x            0.060   0.01    5.10   0.007 **
z           -0.008   0.01   -0.81   0.464
---
Residual standard error: 6.18 on 4 DF
Multiple R-Squared: 0.87, Adj. R-squared: 0.80, cp=3
F-statistic: 13.06 on 2 and 4 DF,  p-value: 0.0176
```

Ο *ηλεκτρονικός υπολογισμός* της γραμμικής παλινδρόμησης

$$y = \beta_0 + \beta_1 z + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

δίνει:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.087   9.944   6.14  0.002 **
```

```
z           -0.003   0.023  -0.13  0.899
---
Residual standard error: 15.14 on 5 DF
Multiple R-Squared: 0.004, Adj. R-squared: -0.20, cp=2
F-statistic: 0.018 on 1 and 5 DF,  p-value: 0.90
```

Ο προσαρμοσμένος συντελεστής προσδιορισμού R^2 λαμβάνει υπόψη τον αριθμό των μεταβλητών και το μέγεθος του δείγματος. Ορίζεται ως:

$$R_a^2 = 1 - \frac{\mathrm{RSS}/(m-n-1)}{\mathrm{TSS}/(m-1)}$$
$$= 1 - (1-R^2)\frac{(m-1)}{(m-n-1)}.$$

Παρατηρήστε ότι R_a^2 μειώνεται όταν η προσθήκη μιας νέας μεταβλητής προκαλέι πολύ μικρή μείωση στο $1-R^2$.

Mallows C_p.

Η διαγραφή μιας εξωγενούς μεταβλητής από το μοντέλο οδηγεί προς την μεροληψία στο μοντέλο. Επιπλέον, η διαγραφή μιας μετεβλητής μειώνει τον πίνακα συνδιακυμάνσεων των εκτιμητών. Το C_p

(έχουντας p ανεξάρτητες) μεταβλητές ορίζεται ως:

$$C_p = \frac{\mathrm{RSS}_p}{\hat{\sigma}^2} - (m - 2p).$$

Αν $C_p \approx p$, τότε το μοντέλο δεν οδηγεί σε μεγάλα επίπεδα μεροληψίας.

Τιμές Σπιτιών – επιλεγμένα μοντέλα

Branch and Bound - exhaustive search

#var.	R^2	Adj.R^2	C_p	Model
0	0.00	-0.04	160.45	const.
1	0.54	0.50	62.56	FLR
2	0.67	0.63	40.04	FLR ST
3	0.76	0.71	26.38	FLR FP ST
4	0.81	0.76	18.94	BDR FLR FP ST
5	0.87	0.82	10.55	BDR FLR FP RMS ST
6	**0.90**	**0.86**	**6.20**	**FLR ST LOT CON GAR L2**
7	0.92	0.88	4.94	BDR FLR ST LOT CON GAR L2
8	0.93	0.89	4.81	BDR FLR RMS ST LOT CON GAR L2
9	0.94	0.89	5.96	BDR FLR FP RMS ST LOT CON GAR L2
10	0.94	0.89	7.51	BDR FLR FP RMS ST LOT BTH CON GAR L2
11	0.94	0.88	9.28	BDR FLR FP RMS ST LOT BTH CON GAR L1 L2
12	0.94	0.87	11.08	BDR FLR FP RMS ST LOT TAX BTH CON GAR
13	0.94	0.86	13.00	BDR FLR FP RMS ST LOT TAX BTH CON GAR

Διαγνωστική παλινδρόμηση

Το 1970 και 80, πολλοί στατιστικολόγοι δημιούργησαν τεχνικές για την εκτίμηση πολλαπλών μοντέλων παλινδρόμησης. Ένα από τα πιο αντιπροσωπευτικά βιβλία στον τομέα ήταν το Regression Diagnostics: Identifying Influential Data and Sources of Collinearity by Belsley, Kuh, and Welch. Ο συγγραφέας Roy Welch αναφέρει πως αρχικά ασχολήθηκε με τη διαγνωστική παλινδρόμηση όταν του ζητήθηκε να εκτιμήσει ένα μοντέλο χρησιμοποιώντας τραπεζικά δεδομένα. Όταν παρουσίασε το εν λόγω μοντέλο στους πελάτες, αυτοι αντέδρασαν αφού το πρόσημο μιας συγκεκριμένης μεταβλητής ήταν αντίθετο από το αναμενόμενο. Το γεγονός αυτό οφειλόταν σε μια ακραία παρατήρηση. Το παράδειγμα αυτό οδήγησε τον Welch στη δημιουργία μεθόδων για την αποφυγή παρόμοιων φαινομένων!

- Σκοπός είναι ο εντοπισμός σημαντικών παρατηρήσεων και μη-σημαντικών μεταβλητών.

- Προβλήματα που σχετίζονται με παρατηρήσεις, δηλαδή, *Ακραίες* και *Επιδρούσες παρατηρήσεις (Influential observations).*

1. Μια παρατήρηση (ή μέτρηση) που είναι ασυνήθιστα μεγάλη ή μικρή σε σχέση με άλλες τιμές ονομάζεται ακραία. Οι ακραίες τιμές μπορεί να οφείλονται σε ένα από τους πιο κάτω λόγους:

   α. Η μέτρηση έχει παρατηρηθεί, καταγραφεί ή εισαχθεί, στον υπολογιστή λανθασμένα.

   β. Οι μετρήσεις προέρχονται από διαφορετικούς πληθυσμούς.

   ς. Η μέτρηση είναι σωστή, αλλά αντιπροσωπεύει ένα σπάνιο φαινόμενο.

2. Επιδρών παρατηρήσεις είναι αυτές που έχουν σημαντική επίδραση στον υπολογισμό της συνάρτησης παλινδρόμησης (δηλαδή, η εκτιμώμενη συνάρτηση παλινδρόμησης διαφέρει σημαντικά με την προσθήκη ή όχι της εν λόγω μεταβλητής στα δεδομένα).

   Με λίγα λόγια, οι επιδρών παρατηρήσεις τραβούν την ευθεία παλινδρόμησης προς το μέρος τους και η διαγραφή τους αλλάζει εντελώς την στατιστική ανάλυση.

• Προβλήματα με τις μεταβλητές. Δηλαδή,

1. Η προσθήκη της δεν βελτιώνει την επεξηγηματικότητα του μοντέλου. Σε αυτή την περίπτωση τεχνικές επιλογής μοντέλου μπορούν να χρησιμοποιηθούν.
2. Μπορεί να είναι όμοια με μια άλλη μεταβλητή (επεξηγούν την ίδια πληροφορία - πολυσυγραμμικότητα (collinearity)). Αυτές οι μεταβλητές πρέπει να εντοπιστούν και/ή να μετασχηματιστεί το μοντέλο, Π.χ. με τη χρήση PCA.
3. Ο αριθμός μεταβλητών μπορεί να έχει παραλειφθεί.

Ο πίνακα HAT

- Δεδομένου του κανονικού μοντέλου παλινδρόμησης:

  $y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_n x_{in}+\varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$

  ή σε συντομογραφία:

  $$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

  Η καλύτερη αμερόληπτη εκτίμηση (BLUE) του β δίνεται από:

  $$\hat{\beta} = (X^T X)^{-1} X^T y.$$

- Οι προβλέψεις του y δίνονται από:

  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_n x_{in}, \quad i = 1, \ldots, m,$

  ή σε μορφή πινάκων:

  $$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= X\big((X^T X)^{-1} X^T y\big) \\ &= X(X^T X)^{-1} X^T y \\ &= Hy \end{aligned}$$

  όπου ο $m \times m$ πίνακας $H = X(X^T X)^{-1} X^T$ ονομάζεται *hat matrix*.

- Ο πίνακας hat είναι *εκθετικά αναλλοίωτος (idempotent)*. Δηλαδή, $H^T = H$ και $H^2 = H$.

- Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του $\hat{y}$ έχει τη μορφή:

$$\mathrm{Var}(\hat{y}) = \begin{pmatrix} \mathrm{Var}(\hat{y}_1) & \mathrm{Cov}(\hat{y}_1, \hat{y}_2) & \dots & \mathrm{Cov}(\hat{y}_1, \hat{y}_m) \\ \mathrm{Cov}(\hat{y}_2, y_1) & \mathrm{Var}(\hat{y}_2) & \dots & \mathrm{Cov}(\hat{y}_2, \hat{y}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{y}_m, \hat{y}_1) & \mathrm{Cov}(\hat{y}_m, \hat{y}_2) & \dots & \mathrm{Var}(\hat{y}_m) \end{pmatrix}$$

- Η διακύμανση-συνδιακύμανση του $\hat{y}$ δίνεται από:

$$\begin{aligned} \mathrm{Var}(\hat{y}) &= \mathrm{Var}(Hy) \\ &= H\mathrm{Var}(y)H^T \\ &= H\sigma^2 I_m H \quad (\text{αφού } \mathrm{Var}(y) = \sigma^2 I_m) \\ &= \sigma^2 H^2 \qquad (\text{αφού } H^T = H \text{ και } I_m H = H) \\ &= \sigma^2 H \qquad (\text{αφού } H^2 = H). \end{aligned}$$

- Τα διαγώνια στοιχεία του H αποτελούν τις διακυμάνσεις του $\hat{y}_i$ για $i = 1, \dots, m$. Δηλαδή,

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 h_{ii}.$$

- Σημειώστε ότι

$$\begin{aligned} h_{11} + h_{22} + \cdots + h_{mm} &= \text{trace}(H) \\ &= \text{trace}(X(X^T X)^{-1} X^T) \\ &= \text{trace}((X^T X)^{-1} X^T X) \\ &= \text{trace}(I_n) \\ &= n. \end{aligned}$$

- Το σύνολο όλων των διακυμάνσεων του $\hat{y}_i$ είναι $n\sigma^2$. Δηλαδή,

$$\sum_{i=1}^{m} \text{Var}(\hat{y}_i) = \text{trace}(\sigma^2 H) = \sigma^2 \sum_{i=1}^{m} h_{ii} = n\sigma^2.$$

- Τα διαγώνια στοιχεία του *hat matrix* h_{ii} ονομάζονται *μοχλεύσεις (leverages)*. Η μόχλευση είναι χρήσιμη στη διαγνωστική στατιστική.

  Σημειώστε ότι η μέση τιμή του h_{ii} είναι n/m. Έτσι, ένας *εμπειρικός κανόνας* προτείνει ότι μόχλευση μεγαλύτερη από $2n/m$ πρέπει να εξεταστεί σε βάθος. Μεγάλες τιμές του h_{ii} οφείλονται σε ακραίες τιμές του X.

Μια παρατήρηση είναι ακραία όταν $h_{ii} > \dfrac{2n}{m}$.

Παράδειγμα (House Data)

Θεωρήστε τον υπολογισμό του μοντέλου:

$$\text{PRICE} = \beta_0 + \beta_1 \text{FLR} + \beta_2 \text{RMS} + \beta_3 \text{BDR} + \beta_4 \text{GAR} + \beta_5 \text{ST} + \varepsilon.$$

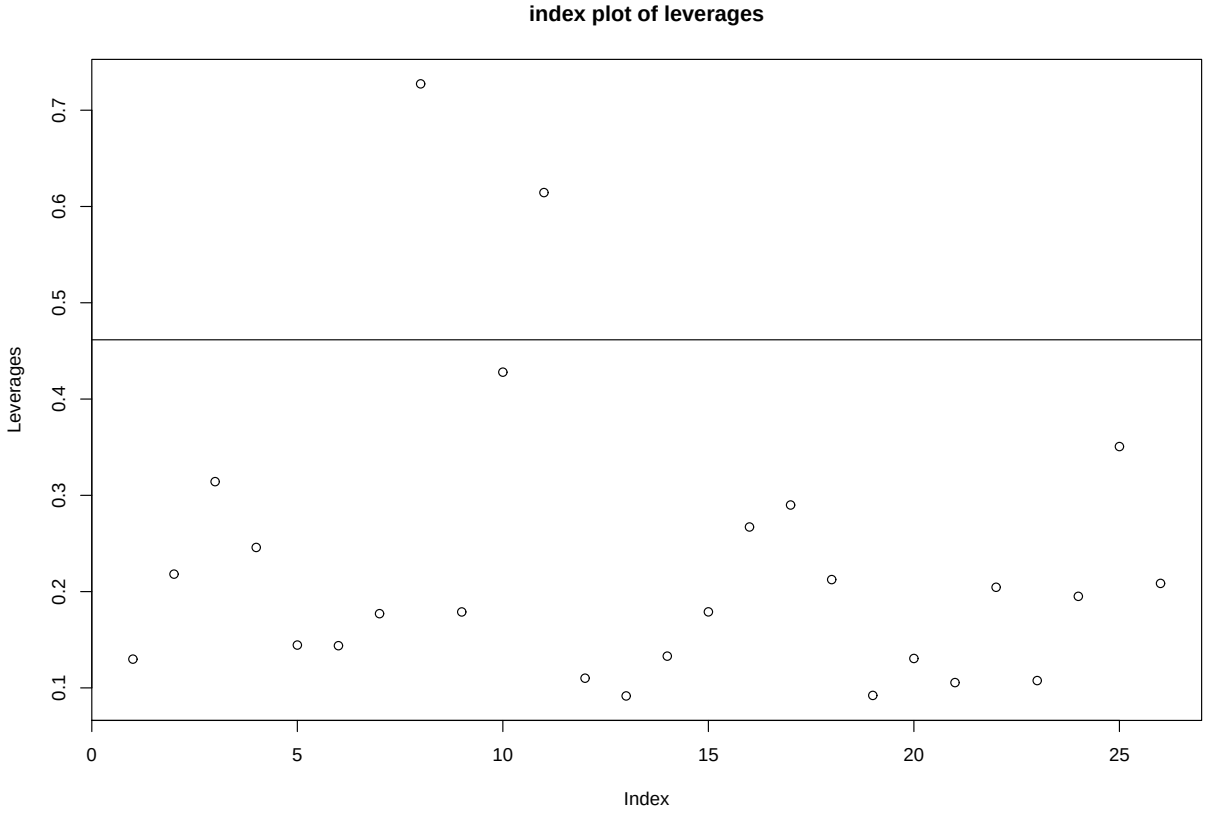
Ο *υπολογιστής* δίνει:

```
Coefficients:
             Estim Std. Error t value Pr(>|t|)
(Intercept) 23.30   5.74   4.06    0.0006 ***
FLR          0.02   0.00   4.15    0.0005 ***
RMS          5.01   1.96   2.55    0.0190 *
BDR         -7.39   2.33  -3.17    0.0049 **
GAR          3.25   1.63   2.00    0.0592 .
ST           9.95   2.77   3.59    0.0018 **
---
Residual standard error: 6.022 on 20 DF
Multiple R-Squared: 0.82, Adjusted R-squared: 0.77
```

Η μόχλευση δίνεται από:

```
leverages <- hat(X)
highlev = 2*6/26 = 0.46          sum(leverages) = 6
> leverages
 0.13 0.22 0.31 0.25 0.14 0.14 0.18 0.73 0.18
 0.43 0.61 0.11 0.09 0.13 0.18 0.27 0.29 0.21
 0.11 0.20 0.11 0.20 0.35 0.21 0.09 0.13
```

```
> leverages[leverages > highlev]
  8   11
0.73 0.61
highlev <- 2*6/26                lowlev <- 1/26.0
plot(leverages, ylab="Leverages",main="index ...)
abline(h=highlev)
```



Με τη διαγραφή της 8ης παρατήρησης έχουμε:

```
Coefficients:
            Estimate Std. Error t  value Pr(>|t|)
(Intercept) 28.22    5.769   4.89  0.0001 ***
FLR          0.029   0.007   4.33 0.00036 ***
RMS          2.076   2.270   0.92 0.37183
BDR         -6.777   2.169  -3.13 0.00558 **
GAR          3.850   1.523   2.53 0.02053 *
ST           9.239   2.576   3.59 0.00199 **
```

```
---
Residual standard error: 5.55 on 19 DF
Multiple R-Squared: 0.84, Adjusted R-squared: 0.80
```

Σφάλματα

- Τα σφάλματα εκφράζονται σε όρους πίνακα ως:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, m.$$

Σε συντομογραφία:

$$\begin{aligned} e &= y - \hat{y} \\ &= y - Hy \\ &= (I_m - H)y. \end{aligned}$$

- Τα σφάλματα αθροίσματος τετραγώνων δίνονται από $\sum_{i=1}^{m} e_i^2$, ή:

$$\begin{aligned} e^T e &= y^T (I_m - H)^T (I_m - H) y \\ &= y^T (I_m - H) y, \end{aligned}$$

καθώς $(I_m - H)$ είναι εκθετικά αναλλοίωτος.
Δηλασή, $(I_m - H) = (I_m - H)^T$ and
$(I_m - H)^2 = (I_m - H)$.

- Η διακύμανση-συνδιακύμανση του e έχει τη μορφή:

$$\mathrm{Var}(e) = \begin{pmatrix} \mathrm{Var}(e_1) & \mathrm{Cov}(e_1, e_2) & \dots & \mathrm{Cov}(e_1, e_m) \\ \mathrm{Cov}(e_2, e_1) & \mathrm{Var}(e_2) & \dots & \mathrm{Cov}(e_2, e_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(e_m, e_1) & \mathrm{Cov}(e_m, e_2) & \dots & \mathrm{Var}(e_m) \end{pmatrix}$$

- Η διακύμανση-συνδιακύμανση του e δίνεται από:

$$\begin{aligned} \mathrm{Var}(e) &= \mathrm{Var}\big((I_m - H)y\big) \\ &= (I_m - H)\mathrm{Var}(y)(I_m - H)^T \\ &= (I_m - H)\sigma^2 I_m (I_m - H) \quad (\text{καθώς } \mathrm{Var}(y) = \sigma^2 I_m) \\ &= \sigma^2 (I_m - H)^2 \\ &= \sigma^2 (I_m - H) \qquad (\text{καθώς } (I_m - H)^2 = (I_m - H)). \end{aligned}$$

- $\mathrm{trace}(I_m - H) = \mathrm{trace}(I_m) - \mathrm{trace}(H) = m - n.$

- Οι διακυμάνσεις του e_i δίνονται από το i διαγώνιο στοιχείο του $\sigma^2(I_m - H)$, δηλαδή,

$$\boxed{\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{για } i = 1, \dots, m.}$$

- Σημειώστε ότι $\mathrm{Var}(e_i) \geq 0$ και έτσι ,

$$\boxed{1 - h_{ii} \geq 0, \quad \text{ή} \quad h_{ii} \leq 1.}$$

Τυποποιημένα σφάλματα (Standardized residuals)

- Θυμηθείτε ότι η διακύμανση των σφαλμάτων $e_i = y_i - \hat{y}_i$ δίνεται από

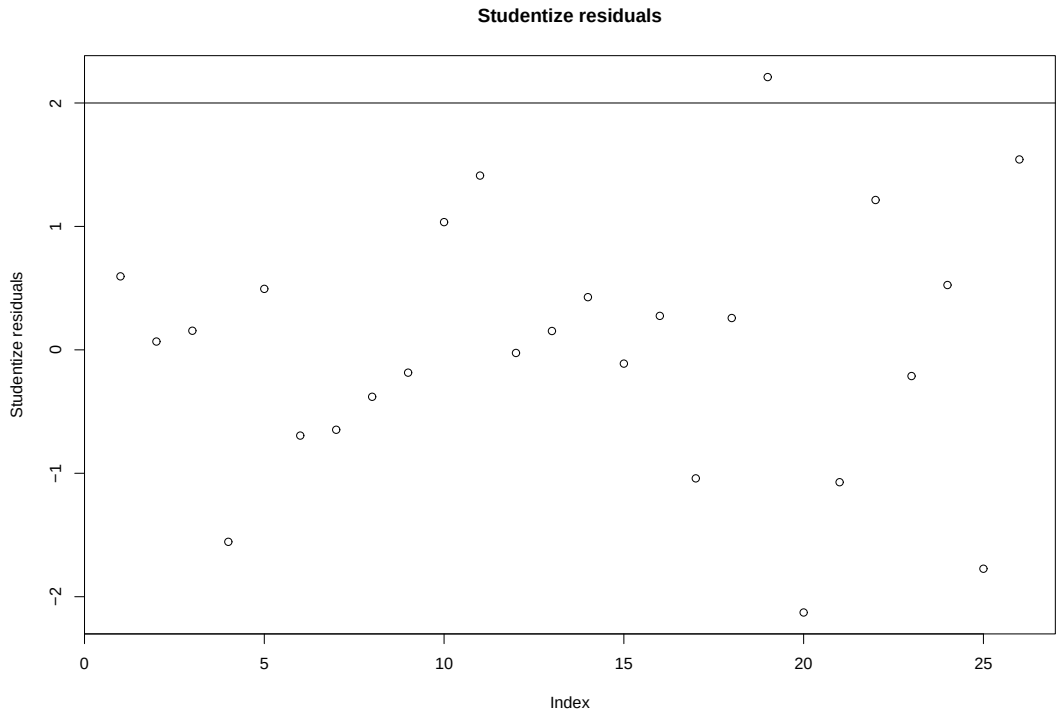
$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{for } i = 1, \ldots, m.$$

- Τα (εσωτερικά) τυποποιημένα σφάλματα δίνονται από:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}.$$

- Αν οι υποθέσεις γραμμικής παλινδρόμησης είναι σωστές, τότε $\mathrm{Var}(r_i) = 1$ και $\mathrm{Cor}(r_i, r_j)$ τείνει να είναι μικρό.

- Υπάρχουν ακραίες τιμές αν $\|r_i\| > 2$.


Studentize residuals
Studentize residuals
Index
2
1
0
−1
−2
0
5
10
15
20
25

Η 26η παρατήρηση των δεδομένων (House data) δίνεται από:

```
Price BDR  FLR FP RMS ST LOT  TAX BTH CON GAR CDN L1 L2
   65   3 1023  0   7  1  30  900 2.0   1 1.0   0  1  0
```

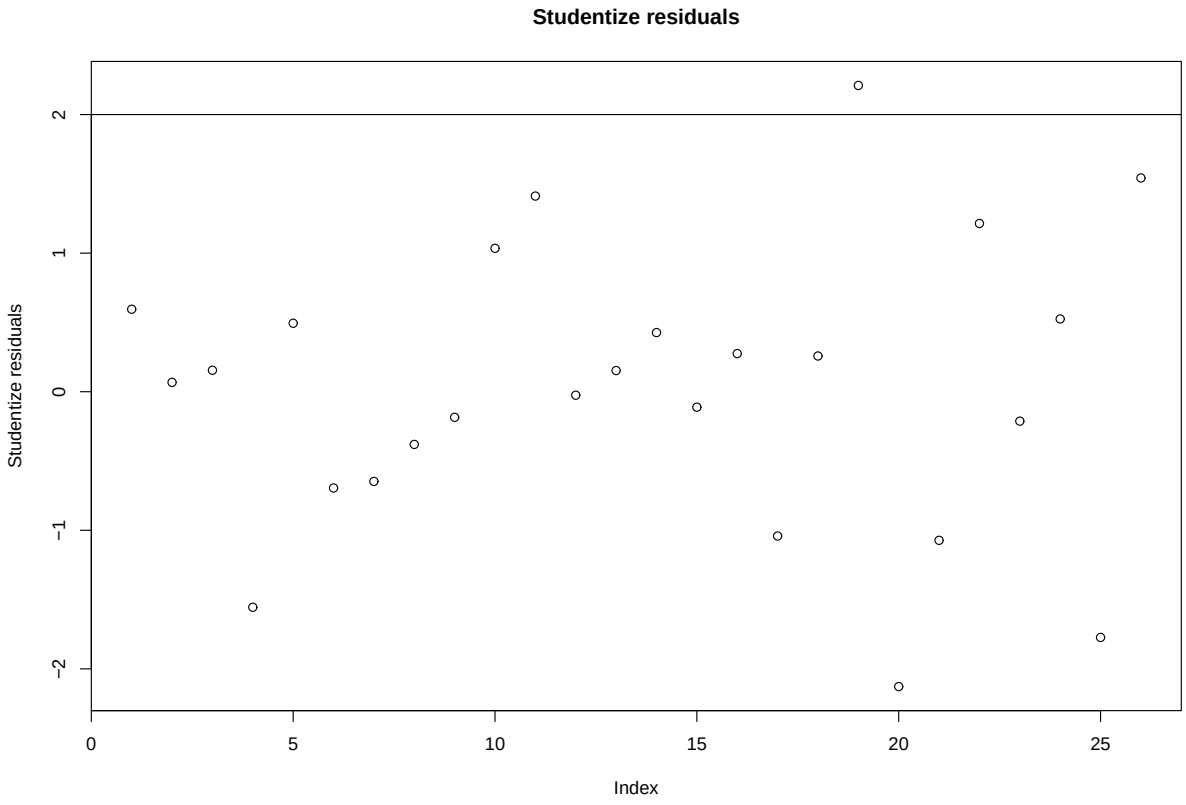
υποθέστε ότι από σφάλμα η τελευταία παρατήρηση αναγράφεται ως:

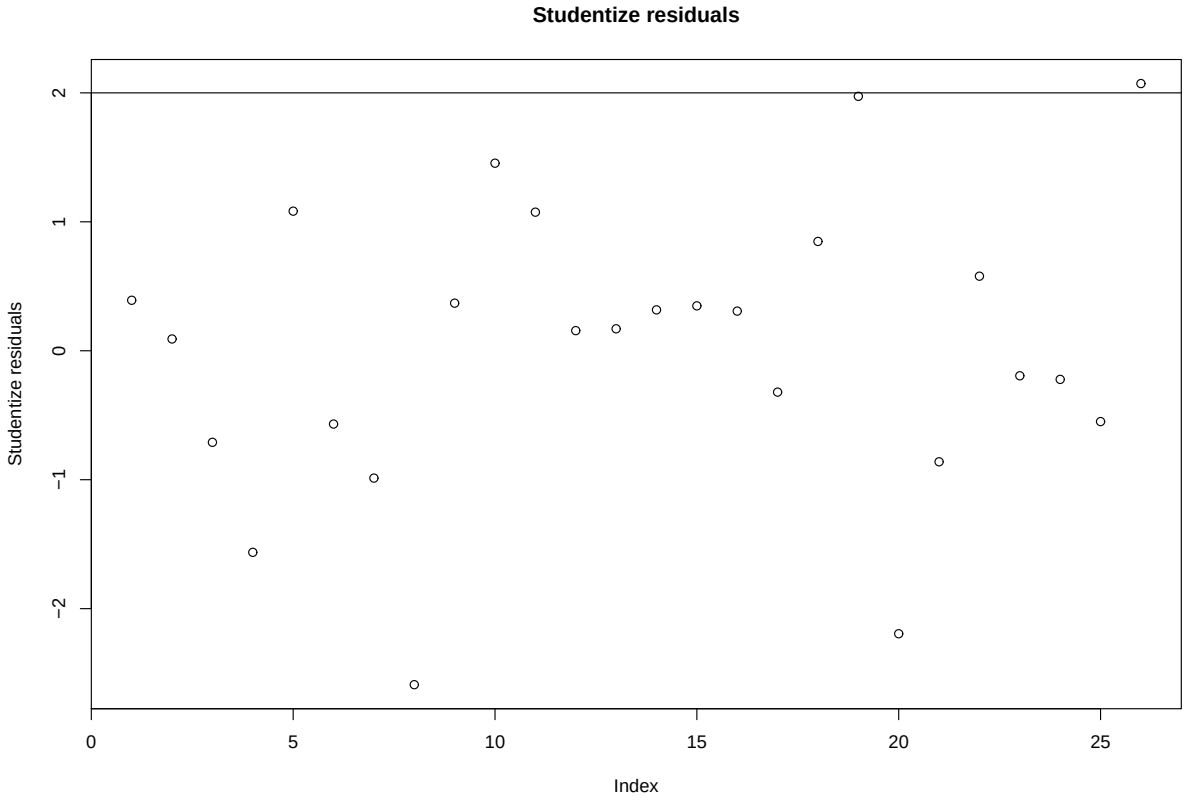
```
Price BDR  FLR FP RMS ST LOT  TAX BTH CON GAR CDN L1 L2
   65   3 1023  0   1  1  30  900 2.0   1 1.0   0  1  0
```

Δηλαδή, RMS αντικαθίστανται με 1 (αντί για 7).

Οι νέοι εκτιμητές δίνονται από:

```
Coefficients:
            Estimate Std.  Error t value Pr(>|t|)
(Intercept) 32.69    4.57    7.16 6.21e-07 ***
FLR          0.02    0.01    4.56 0.000190 ***
RMS          1.18    1.21    0.97 0.342454
BDR         -3.70    1.96   -1.89 0.073696 .
GAR          3.49    1.83    1.91 0.070673 .
ST          11.43    3.27    3.50 0.002263 **
---
Residual standard error: 6.78 on 20 DF
Multiple R-Squared: 0.77, Adjusted R-squared: 0.71
```


Studentize residuals
Studentize residuals
Index
−2
−1
0
1
2
0
5
10
15
20
25


Studentize residuals
Studentize residuals
Index
−2
−1
0
1
2
0
5
10
15
20
25

Επιδρούσες Παρατηρήσεις: Απόσταση Cook's

- Επιδρών σημείο είναι αυτό του οποίου η αφαίρεση προκαλεί μεγάλη αλλαγή στις εκτιμήσεις (υπολογισμό του μοντέλου). Ένα επιδρών σημείο μπορεί να είναι ή όχι ακραία τιμή και μπορεί να έχει ή όχι μεγάλη μόχλευση. Σίγουρα όμως θα έχει ένα από τα δύο προηγούμενα χαρακτηριστικά.

- Θώτουμε τον κάτω δείκτη i να υποδηλώνει το υπολογισμό (εκτίμηση) χωρίς την παρατήρηση (i). Πιο κάτω δίνονται τρόποι μέτρησης της επιρροής αυτής της παρατήρησης:

  1. Αλλαγή στους συντελεστές: $\hat{\beta} - \hat{\beta}_{(i)}$.

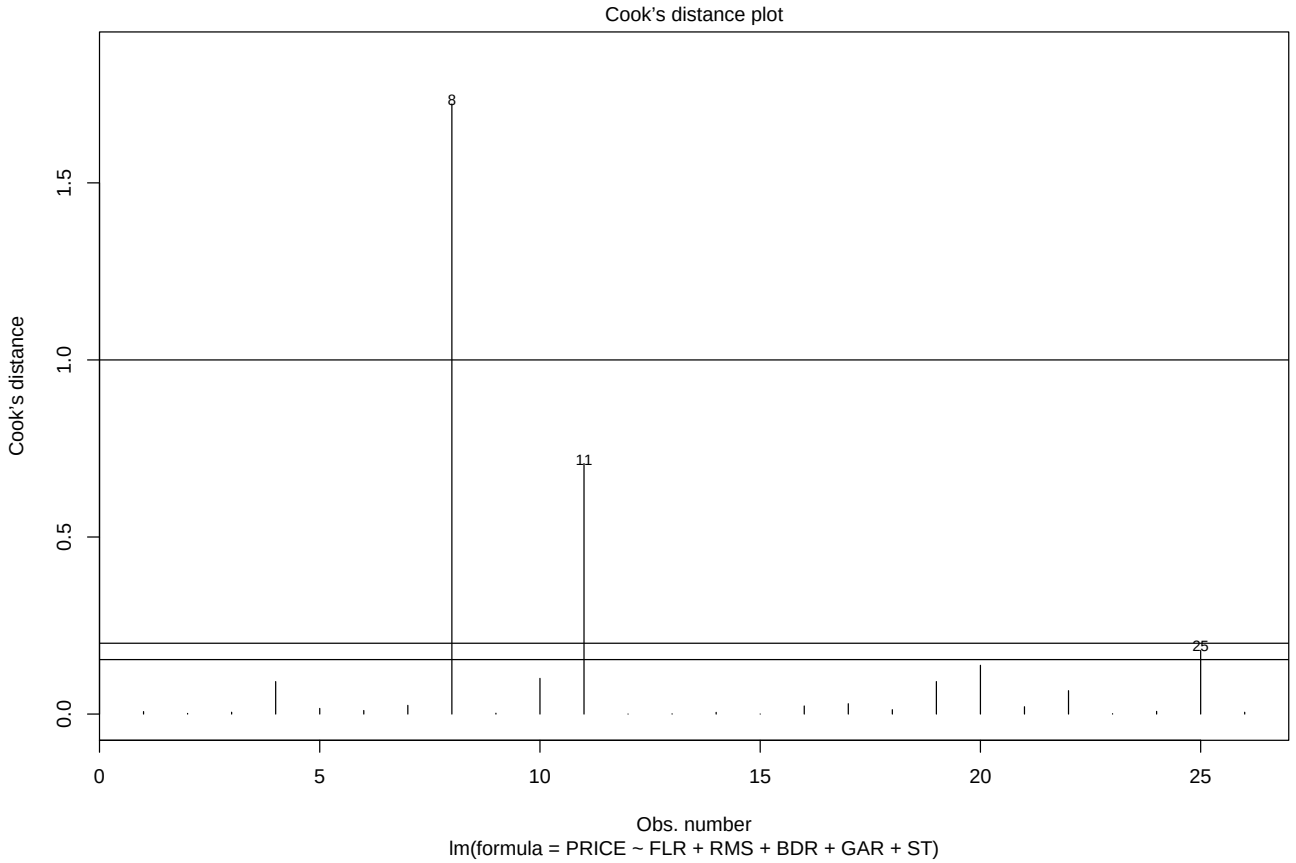
  2. Αλλαγή στην εκτίμηση:
     $\hat{y} - \hat{y}_{(i)} = X^T(\hat{\beta} - \hat{\beta}_{(i)})$.

- Με τους πιο τρόπους η κλίμακα μπορεί να διαφέρει μεταξύ διαφορετικών δειγμάτων. Δεν υπάρχει όμοιο μέτρο σύγκρισης των αποτελεσμάτων. Μια

δημοφιλής εναλλακτική είναι η απόσταση Cook's:

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{n\hat{\sigma}^2}$$
$$= \frac{r_i^2}{n} \frac{h_{ii}}{1 - h_{ii}}.$$

- Η απόσταση Cook's, D_i, μετρά την επιρροή ενός στοιχείου. Η απόσταση Cook's μετρά την επίδραση της διαγραφής μιας συγκεκριμένης μεταβλητής. Παρατηρήσεις με μεγαλύτερες τιμές D_i απότι αυτές των άλλων παρατηρήσεων του δείγματος θεωρούνται ότι έχουν ασυνήθιστη μόχλευση. Σημείο αποκοπής για τον εντοπισμό επιδρουσών παρατηρήσεων θεωρείτε το D_i μεγαλύτερο του $4/(m-n)$, όπου m είναι ο αριθμός των παρατηρήσεων και n είναι ο αριθμός των ανεξάρτητων μεταβλητών (συμπεριλαμβανομένης και της σταθεράς). Άλλοι προτείνουν το $D_i > 1$ σαν δυνατή ένδειξη ύπαρξης ακραίων παρατηρήσεων, με $D_i > 4/m$ το κριτήριο ένδειξης πιθανού προβλήματος.



Συγγραμικότητα (Collinearity)

- Ο βαθμός στον οποίο οι ανεξάρτητες μεταβλητές συσχετίζονται, και κατά συνέπεια προβλέπουν η μια την άλλη, ονομάζεται συγγραμικότητα. Αν οι συγγραμικότητα κάποιων ανεξάρτητων μεταβλητών είναι τόσο μεγάλη ώστε οι ανεξάρτητες μεταβλητές να προβλέπουν άλλες ανεξάρτητες μεταβλητές τότε το φαινόμενο αυτό ονομάζεται πολυσυγγραμικότητα (multicollinearity).

- Η πολυσυγγραμικότητα προκαλέι προβλήματα

εξαγωγής συμπερασμάτων σχετικά με τη σχέση μεταξύ εκτιμητών και εξαγώμενης μεταβλητής με τη χρήση των μοντέλων παλινδρόμησης. Η p-τιμή ενός εκτιμητή μπορεί να υποδηλώνει μη-σημαντικότητα της συγκεκριμένη μεταβλητής, ενώ στην ουσία να είναι σημαντική. Τα διαστήματα εμπιστοσύνης για τους συντελεστές παλινδρόμησης σε ένα πολυσυγγραμικό μοντέλο, μπορεί να είναι τόσο μεγάλα ώστε έστω και μια πολύ μικρή αλλαγή στις παρατηρήσεις να έχει μεγάλη επίδραση στους συντελεστές. Κάποιες φορές μπορεί να αλλάζει και το πρόσημο του συντελεστή.

- Μία προφανής μέθοδος αξιολόγησης του βαθμού συσχέτισης μιας ανεξάρτητης μεταβλητής με άλλες είναι η εξέταση του R_j^2. Αυτή είναι η τιμή του συντελεστή προσδιορισμού R^2 μεταξύ της μεταβλητής x_j και όλων των άλλων ανεξάρτητων μεταβλητών. Δηλαδή, R_j^2 είναι το R^2 που θα πάρουμε αν παλινδρομήσουμε το x_j με όλα τα άλλα x_i΄ς.

- Ο παράγοντας αντοχής TOL_j ορίζεται ως:

$$\mathrm{TOL}_j = 1 - R_j^2.$$

- TOL_j είναι κοντά στο 1 αν x_j δεν είναι στενά συσχετισμένο με τους άλλους εκτιμητές.

- Ο παράγοντας πληθωριστικής διακύμανσης, Variance Inflation Factor (VIF) (και το αντίστροφο, η ανοχή (tolerance), ως μέτρο πολυγγραμικότητας:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}.$$

- Μία τιμή του VIF_i κοντά στο 1 υποδηλώνει μη συσχέτιση, ενώ μεγαλύτερες τιμές υποδειλώνουν την παρουσία πολυσυγγραμικότητας (αμελητεές πληροφορίες στις επεξηγηματικές μεταβλητές).

  Π.χ. αν $R_j^2 = 0.90$, τότε $\mathrm{VIF}_i = 10$ και συνιστάται προσοχή (κάποιου άλλοι θεωρούν $\mathrm{VIF}_i = 5$, δηλαδή, $R_j^2 = 0.80$).

- Ο πίνακας συσχετίσεως των ανεξάρτητων μεταβλητών, έστω R, μπορεί επίσης να χρησιμοποιηθεί για τον εντοπισμό πολυσυγγραμικότητας. Όμως οι δυσκολία είναι ότι το R υποδηλώνει τη σχέση μεταξύ μεμονωμένων ζευγαριών μεταβλητών και έτσι δεν μπορεί να

ανιχνεύσει τις σχέσεις μεταξύ κάθε x_j και όλων των άλλων εκτιμητών. Εντούτοις, τα i διαγώνια στοιχεία του R^{-1} είναι ο VIF_i.

- Ο *αριθμός κατάστασης (condition number)* του εξωγενούς πίνακα X μπορεί να μας ενημερώσει για τη γραμμική συσχέτιση μεταξύ των εξωγενών μεταβλητών. Δηλαδή,

$$\eta = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \geq 1,$$

όπου λ_j $(j = 1, \ldots, n)$ είναι οι ιδιοτιμές του X.

- Γενικά ο *αριθμός κατάστασης*

$$\eta_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}, \quad j = 1, \ldots, n$$

υποδηλώνει μέτρια έως δυνατή σχέση αν $\eta_j > 30$.

Επιπλέον

Ο BLUE του τυπικού γραμμικού μοντέλου παλινδρόμησης $y = X\beta + \varepsilon$ *δίνεται από:*

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Αν υπάρχει πολυσυγγραμικότητα, τότε ο αριθμός κατάστασης, η, είναι ψηλός. Δηλαδή, ο $(X^T X)^{-1}$ είναι ακριβώς ή κατά προσέγγιση ιδιάζων. Έτσι, δεν θα υπάρχει λύση ή οι εκτιμητές θα είναι άχρηστοι.

Παράδειγμα

Θεωρήστε τις ψηλές πολυσυγγραμικές τιμές των ανεξάρτητων μεταβλητών x_1 και x_2 που δίνονται στον παρακάτω πίνακα. Οι εξαρτημένες μεταβλητές $y^{(1)}$, $y^{(2)}$ και $y^{(3)}$ προέρχονται από διαφορετικά δείγματα. Δημιουργήθηκαν με την προσθήκη ενός τυχαίου αριθμού από κανονική κατανομή $N(0, 0.01)$ στην:

$$x_1 + 2\,x_2$$

και οι αντίστοιχες τιμές των εξαρτημένων αυτών μεταβλητών είναι πολύ όμοιες.

x_1	x_2	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
2.705	2.695	8.12	8.09	8.09
2.995	3.005	9.01	9.02	9.00
3.255	3.245	9.74	9.75	9.74
3.595	3.605	10.82	10.80	10.79
3.805	3.795	11.38	11.39	11.40
4.145	4.155	12.44	12.44	12.45
4.405	4.395	13.19	13.20	13.19
4.745	4.755	14.27	14.25	14.25
4.905	4.895	14.68	14.70	14.71
4.845	4.855	14.56	14.55	14.54

Ο VIF του x_1 και x_2 δίνεται από 5868.7.

Ο αριθμός κατάστασης του $(x_1\ x_2)$ είναι 802.7.

Για το $y^{(1)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   0.5926     0.4160   1.425 0.192068
x2   2.4070     0.4159   5.787 0.000411 ***
---
Signif. codes:
Residual standard error: 0.013 on 8 DF
Multiple R-Squared: 1,  Adjusted R-squared:1
F-statistic: 4.19e+06 on 2 and 8 DF, p-value:< 2.2e-16
```

Για το $y^{(2)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
```

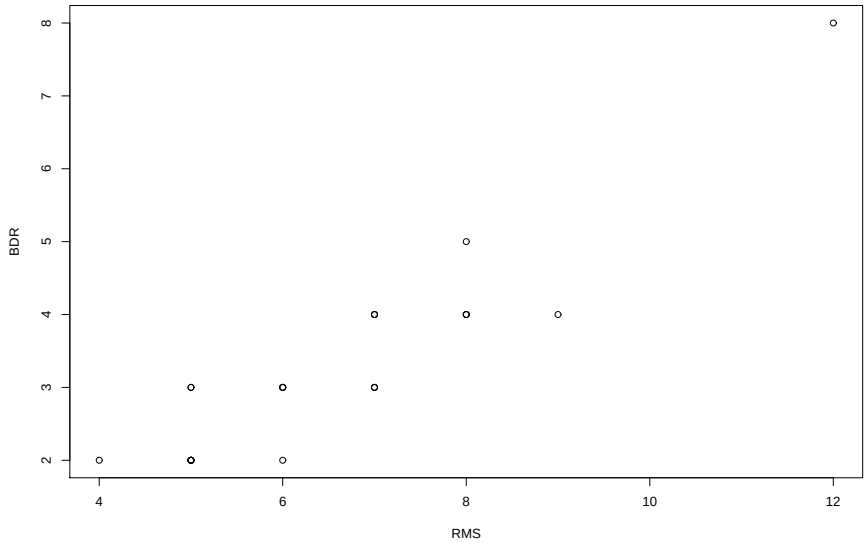
```
x1   1.20     0.28     4.27   0.0027 **
x2   1.80     0.28     6.39   0.0002 ***
---
Residual standard error: 0.0089 on 8 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 9.14e+06 on 2 and 8 DF, p-value: < 2.2e-16
```

Για το $y^{(3)} = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   1.46     0.26   5.71 0.0004 ***
x2   1.54     0.26   6.05 0.0003 ***
---
Residual standard error: 0.008 on 8 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 1.1e+07 on 2 and 8 DF,  p-value: < 2.2e-16
```

Παράδειγμα (*Μοντέλο House prices*)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.30   5.74   4.06 0.0006 ***
FLR          0.02   0.01   4.15 0.0005 ***
RMS          5.01   1.96   2.55 0.0190 *
BDR         -7.39   2.33  -3.17 0.0049 **
GAR          3.25   1.63   2.00 0.0592 .
ST           9.95   2.77   3.59 0.0018 **
---
Residual standard error: 6.02 on 20 DF
Multiple R-Squared: 0.82, Adjusted R-squared: 0.77
F-statistic: 17.82 on 5 and 20 DF, p-value: 9.2e-07
```



Οι VIF δίνονται από:

```
vif(House)
 FLR  RMS  BDR  GAR   ST
2.43 7.70 6.40 1.23 1.08
```

Το επίπεδο πολυσυγγραμικότητας των RMS και BDR θεωρείται αρκετά ψηλό.

Υπολογισμός VIF στο μοντέλο *House prices*

- Υπολογισμός παλινδρομικού μοντέλου:

$$\mathrm{FLR} = \beta_0 + \beta_1 \mathrm{RMS} + \beta_2 \mathrm{BDR} + \beta_3 \mathrm{GAR} + \beta_4 \mathrm{ST} + \varepsilon.$$

Αυτό δίνει $R^2_{\mathrm{FLR}} = 0.589$ και κατά συνέπεια:

$$\mathrm{VIF}_{\mathrm{FLR}} = 1/(1 - R^2_{\mathrm{FLR}}) = 2.43.$$

- Υπολογισμός παλινδρομικού μοντέλου:

$$\mathrm{RMS} = \beta_0 + \beta_1 \mathrm{FLR} + \beta_2 \mathrm{BDR} + \beta_3 \mathrm{GAR} + \beta_4 \mathrm{ST} + \varepsilon.$$

δίνει $R^2_{\mathrm{RMS}} = 0.87$ και $\mathrm{VIF}_{\mathrm{RMS}} = 1/(1 - 0.87) = 7.69$.

- Υπολογισμός παλινδρομικού μοντέλου:

$$\mathrm{BDR} = \beta_0 + \beta_1 \mathrm{FLR} + \beta_2 \mathrm{RMS} + \beta_3 \mathrm{GAR} + \beta_4 \mathrm{ST} + \varepsilon.$$

δίνει $R^2_{\mathrm{BDR}} = 0.84$ και $\mathrm{VIF}_{\mathrm{BDR}} = 1/(1 - 0.84) = 6.25$.

- Υπολογισμός παλινδρομικού μοντέλου:

$$\mathrm{GAR} = \beta_0 + \beta_1 \mathrm{FLR} + \beta_2 \mathrm{RMS} + \beta_3 \mathrm{BDR} + \beta_4 \mathrm{ST} + \varepsilon.$$

δίνει $R^2_{\mathrm{GAR}} = 0.19$ και $\mathrm{VIF}_{\mathrm{GAR}} = 1/(1 - 0.19) = 1.23$.

- Υπολογισμός παλινδρομικού μοντέλου:

$$\mathrm{ST} = \beta_0 + \beta_1 \mathrm{FLR} + \beta_2 \mathrm{RMS} + \beta_3 \mathrm{BDR} + \beta_4 \mathrm{GAR} + \varepsilon.$$

δίνει $R^2_{\mathrm{ST}} = 0.08$ και $\mathrm{VIF}_{\mathrm{ST}} = 1/(1 - 0.08) = 1.08$.

Υπολογισμός VIF από πίνακα συσχέτισης

- Ο πίνακας συσχέτισης, έστω R, των ανεξάρτητων μεταβλητών δίνεται από:

	FLR	RMS	BDR	GAR	ST
FLR	1.00	0.74	0.68	0.40	0.13
RMS	0.74	1.00	0.92	0.30	0.23
BDR	0.68	0.92	1.00	0.24	0.23
GAR	0.40	0.30	0.24	1.00	0.17
ST	0.13	0.23	0.23	0.17	1.00

- Ο αντίστροφος του πίνακα συσχέτισης, δηλαδή R^{-1} δίνεται από:

	FLR	RMS	BDR	GAR	ST
FLR	2.43	-1.62	-0.07	-0.51	0.17
RMS	-1.62	7.70	-5.87	-0.21	-0.21
BDR	-0.07	-5.87	6.40	0.28	-0.13
GAR	-0.51	-0.21	0.28	1.23	-0.16
ST	0.17	-0.21	-0.13	-0.16	1.08

- Παρατηρήστε ότι τα διαγώνια στοιχεία του R^{-1} είναι οι VIF των ανεξάρτητων μεταβλητών. Δηλαδη, $\text{Diag}(R^{-1}) = (2.43, \quad 7.70, \quad 6.40, \quad 1.23, \quad 1.08) \equiv \text{VIF}$.

- Αν δεν υπάρχει πολυσυγγραμικότητα, τότε R, και κατά συνέπεια R^{-1}, έχουν 1 στα διαγώνια στοιχεία και μηδέν στο υπόλοιπα. Οι VIF δείχνουν σε πιο βαθμό η διακύμανση μιας μεμονωμένης μεταβλητής

έχει επιρεαστεί από την παρουσία πολυσυγγραμικότητας.

Σύνοψη και παράδειγμα (διαγνωστική παλινδρ.)

- Υποθέτουμε ότι υπάρχει γραμμική συσχέτιση μεταξύ των χρόνων εκπαίδευσης (EDU), ηλικία (AGE) και μισθό (SAL). Θεωρήστε το μοντέλο παλινδρόμησης:

$$\mathrm{SAL}_i = \beta_0 + \beta_1 \mathrm{EDU}_i + \beta_2 \mathrm{AGE}_i + \varepsilon_i.$$

- Τα δεδομένα που χρησιμοποιηθήκαν στο μοντέλο δίνονται πιο κάτω:

SAL $K	EDU χρόνια	AGE χρόνια
26.2	12	34
46.5	9	40
28.6	15	37
28.8	16	36
30.4	18	38
34.2	22	44
34.9	24	43

- Το διάνυσμα συντελεστών β και δεδομένων y και ο πίνακας X στο μοντέλο παλινδρόμησης $y = X\beta + \varepsilon$

δίνονται από:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, y = \begin{matrix} \text{SAL} \\ \begin{pmatrix} 26.2 \\ 46.5 \\ 28.6 \\ 28.8 \\ 30.4 \\ 34.2 \\ 34.9 \end{pmatrix} \end{matrix}, \quad X = \begin{matrix} (1 & \text{EDU} & \text{AGE}) \\ \begin{pmatrix} 1 & 12 & 34 \\ 1 & 9 & 40 \\ 1 & 15 & 37 \\ 1 & 16 & 36 \\ 1 & 18 & 38 \\ 1 & 22 & 44 \\ 1 & 24 & 43 \end{pmatrix} \end{matrix}.$$

- Ο πίνακας HAT δίνεται από: $H = X(X^T X)^{-1} X^T$.

$$H = \begin{pmatrix} 0.43 & 0.08 & 0.25 & 0.31 & 0.19 & -0.17 & -0.11 \\ 0.08 & 0.93 & 0.10 & -0.08 & -0.07 & 0.16 & -0.12 \\ 0.25 & 0.10 & 0.19 & 0.21 & 0.17 & 0.03 & 0.05 \\ 0.31 & -0.08 & 0.21 & 0.29 & 0.22 & -0.03 & 0.07 \\ 0.19 & -0.07 & 0.17 & 0.22 & 0.20 & 0.10 & 0.18 \\ -0.17 & 0.16 & 0.03 & -0.03 & 0.10 & 0.47 & 0.43 \\ -0.11 & -0.12 & 0.05 & 0.07 & 0.18 & 0.43 & 0.49 \end{pmatrix}$$

- Τα διαγώνια στοιχεία του H πίνακα, δηλαδή, h_{ii} $(i = 1, \ldots, m)$, δείχνουν τη μόχλευση του μοντέλου: leverages $= (0.43,\ 0.93,\ 0.19,\ 0.29,\ 0.20,\ 0.47,\ 0.49)$.

- Η διακύμανση των εκτιμώμενων τιμών $\hat{y}_i$ δίνονται

από $\sigma^2 h_{ii}$.

- Το άθροισμα των μοχλεύσεων (δηλαδή, το άθροισμα των διαγώνιων στοιχείων του $H \equiv \sum_{i=1}^{m} h_{ii}$) δίνεται από τον αριθμό των μεταβλητών στο μοντέλο (συμπεριλαμβανομένης και της αποτέμνουσας).

$$\sum_{i=1}^{m} h_{ii} = 0.43+0.93+0.19+0.29+0.20+0.47+0.49 = 3.$$

- Αυτό υπονοεί ότι το σύνολο των διακυμάνσεων των εκτιμημένων τιμων του $\hat{y}_i$ ισούνται με τον αριθμό των μεταβλητών στο μοντέλο πολλαπλασιασμένο με σ^2:

$$\sum_{i=1}^{m} \mathrm{Var}(\hat{y}_i) = \sigma^2 \sum_{i=1}^{m} h_{ii} = n\sigma^2.$$

- Αν όλες οι διακυμάνσεις των εκτιμώμενων τιμών είναι όμοιες τότε:

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 h_{ii} = \sigma^2 \frac{n}{m}, \quad \text{or} \quad h_{ii} = \frac{n}{m}.$$

Στο παράδειγμα $m = 7$ και $n = 3$. Έτσι, $n/m = 0.429$.

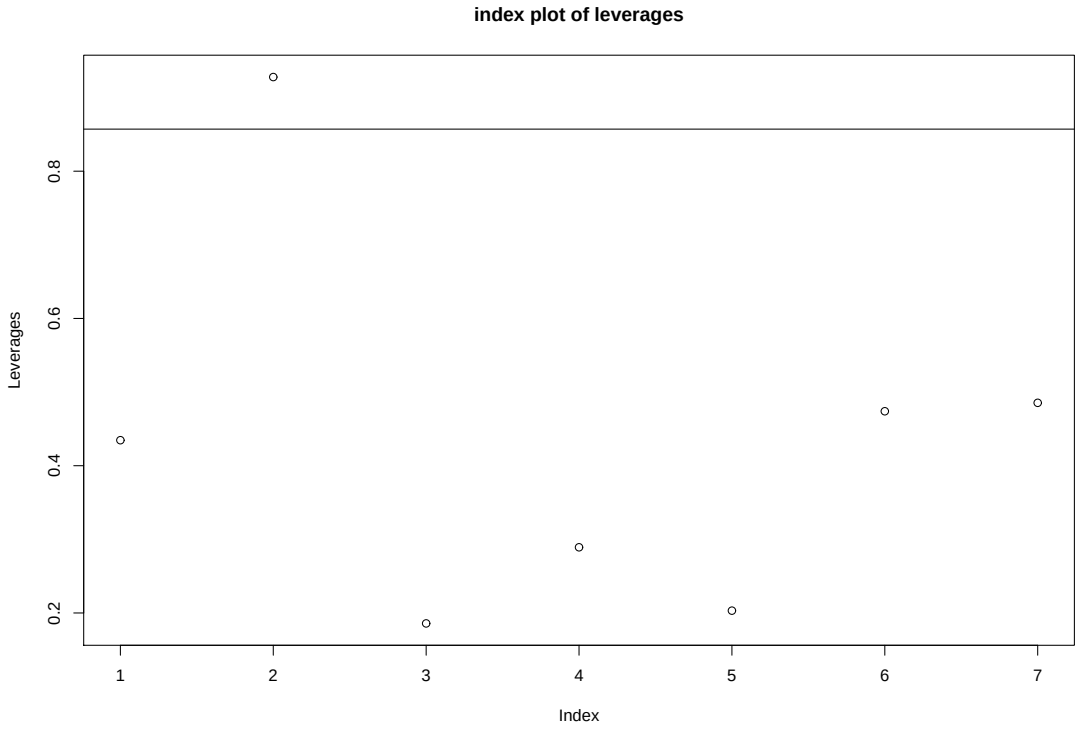
- Μια παρατήρηση είναι επιδρών αν η εκτιμώμενη της τιμή έχει *ίση* μεγαλύτερη διακύμανση από τη μέση διακύμανση. Εδώ, διπλάσια δηλώνει μεγάλη διακύμανση. Δηλαδή,

$$\textit{The } i\textit{th observation is influential if} \quad h_{ii} > \frac{2n}{m}.$$

- Στο παράδειγμα των (μισθών) μια παρατήρηση με μόχλευση μεγαλύτερη του $2 \times 0.429 = 0.86$ είναι επιδρών.

- Ένα γράφημα μπορεί επίσης να χρησιμοποιηθεί στον εντοπισμό επιδρών παρατηρήσεων.

```
            Estimate   SE   t value Pr(>|t|)
(Intercept) -33.27  12.83  -2.59  0.061 .
edu          -1.28   0.27  -4.70  0.009 **
age           2.25   0.39   5.72  0.005 **
---
Residual standard error: 2.69 on 4 DF
Multiple R-Squared: 0.90, Adjusted R-squared: 0.84
F-statistic: 17.2 on 2 and 4 DF, p-value: 0.011
```



- Η δεύτερη παρατήρηση είναι επιδρούσα. Η διαγραφή της παρατήρησης αλλάζει το εκτιμώμενο μοντέλο ως εξής:

```
            Estimate  SE          t value     Pr(>|t|)
(Intercept) 10    3.439e-14  2.908e+14   <2e-16 ***
edu         0.5   1.272e-15  3.929e+14   <2e-16 ***
age         0.3   1.435e-15  2.091e+14   <2e-16 ***
---
Residual standard error: 3.67e-15 on 3 DF
```

- Τα σφάλματα δίνονται από:

$$e_i = y_i - \hat{y}_i$$
$$= y_i - \hat{\beta}_0 + \hat{\beta}_1 \mathrm{EDU}_i + \hat{\beta}_2 \mathrm{AGE}_i, \quad \text{for } i = 1, \ldots, m.$$

- Στο παράδειγμα των *Μισθών* τα σφάλματα είναι:

$$e = (-1.54 \quad 1.44 \quad -2.04 \quad 1.69 \quad 1.35 \quad -3.20 \quad 2.30)^T$$

- Η διακύμανση του e_i δίνεται από

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{για } i = 1, \ldots, m.$$

- Η διακύμανση είναι θετική. Έτσι, $(1 - h_{ii}) > 0$ το οποίο υπονοεί ότι

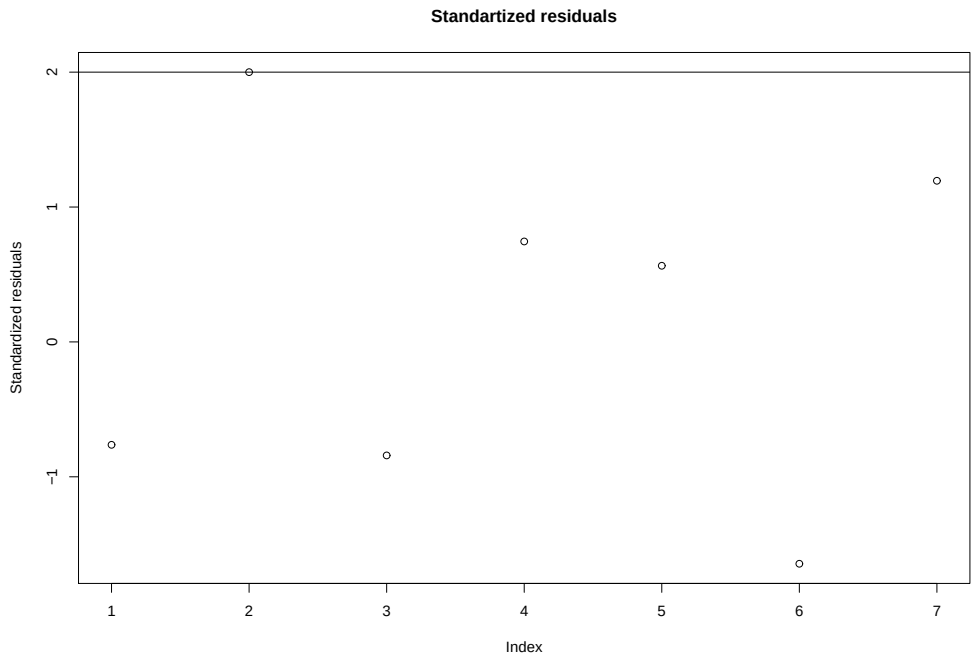
$$0 \leq h_{ii} \leq 1.$$

- Τα τυποποιημένα σφάλματα (standardized residuals) δίνονται από:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}.$$

- Μια παρατήρηση είναι ακραία αν

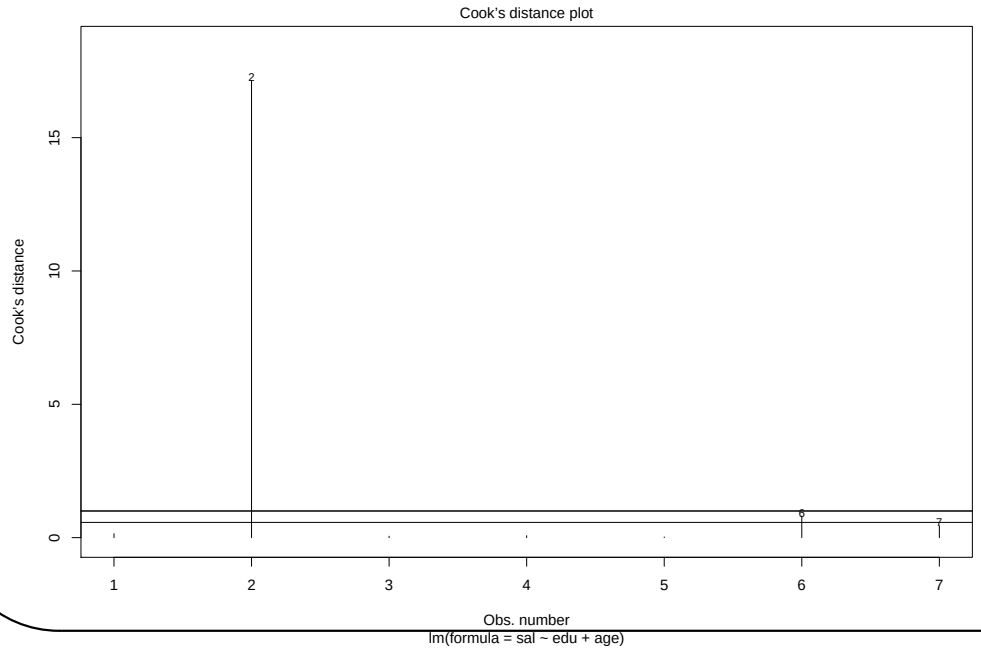
$$\|r_i\| > 2.$$



- Η απόσταση Cook's, ορίζεται από D_i, είναι ακόμη ένα μέτρο εντοπισμού επιδρών σημείων:

- Το σημείο αποκοπής για τον εντοπισμό επιδρών παρατηρήσεων είναι:

$$D_i = 1, \quad \text{ή} \quad D_i > 4/m, \quad \text{ή} \quad D_i > 4/m.$$



Περιορισμένη παλινδρόμηση ελαχ. τετραγ.

Συχνά απαιτείται η εισαγωγή περιορισμών στην εκτίμηση του OLS. Αυτή η περίπτωση εκτίμησης OLS ονομάζεται περιορισμένη ελαχίστων τετραγώνων, restricted least squares (RLS). Θεωρήστε το κανονικό μοντέλο γραμμικής παλινδρόμησης

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_m).$$

Σε κάποιες περιπτώσεις πρέπει να ορίσουμε εκ των προτέρων περιορισμούς που επιρεάζουν το διάνυσμα β που έχουν τη μορφή:

$$R\beta = r,$$

όπου R είναι ένας $j \times n$ πίνακας με $j < n$ και $\text{rank}(R) = j$. Η πληροφορία θεωρείται αληθής και οι εξισώσεις είναι αλγεβρικά συνεπείς.

Η κατάσταση $\text{rank}(R) = j$ υπονοεί ότι οι γραμμές του R είναι γραμμικά ανεξάρτητες και έτσι, κανένας περιορισμός δεν είναι αμελητέος.

Παράδειγμα

Η συνάρτηση παραγωγής Cobb-Douglas δίνεται από:

$$Q = \alpha\, C^{\gamma} L^{(1-\gamma)},$$

όπου Q είναι η παραγωγή (απόδοση), C είναι η ποσότητα του κεφαλαίου, L είναι η ποσότητα του εργατικού δυναμικού και α είναι μια σταθερά[α΄].

Για την εκτίμηση της συνάρτησης μετασχηματίζουμε με τη χρήση του λογαρίθμου, logs, και προσθέτουμε ένα σφάλμα:

$$\begin{array}{lll} \log Q = \log \alpha + \gamma\, \log C & + (1-\gamma)\, \log L + \varepsilon \\ y \qquad = \beta_0 \quad + \beta_1\, \log C + \beta_2\, \log L & + \varepsilon. \end{array}$$

Τότε $\beta_1 = \gamma$ και $\beta_2 = 1 - \gamma$ είναι συντελεστές που υπόκεινται στους περιορισμούς ότι $\beta_1 + \beta_2 = 1$.
Δηλαδή, $R\beta = r$ υπονοεί $R = (0\ 1\ 1)$,
$\beta = (\beta_0\ \beta_1\ \beta_2)^T$ και $r = 1$.

[α΄] Είναι μια θετικά ορισμένη σταθερά που υποδειλώνει την ποσότητα της παραγωγής που 1 μονάδα εργατικού δυναμικού και μία μονάδα καφαλαίου μπορούν να παράξουν. Η σταθερά γ δειλώνει ότι μια σχετική αύξηση στο Q ισοδυναμεί σε μία ιση σχετική αύξηση στον C και L. Έτσι, αν C και L διπλασιαστούν, τότε Q διπλασιάζεται και έτσι, υπάρχουν σταθερές οικονομίες κλίμακας.

Μπορούμε να εξάγουμε τα περιορισμένα ελάχιστα τετράγωνα για την εκτίμηση των εξισώσεων με την αξιολόγηση της ακόλουθης συνάρτησης Lagrangian:

$$L = (y - X\beta)^T(y - X\beta) + 2\lambda(R\beta - r)$$

Παραγοντοποιούμε ως προς το β και θέτουμε το αποτέλεσμα ίσο με μηδέν. Έτσι έχουμε:

$$\frac{dL}{d\beta} = -2y^T X + 2\beta^T X^T X + 2\lambda R = 0$$

από το οποίο παίρνουμε

$$X^T X\beta + R^T\lambda = X^T y.$$

Πολλαπλασιάζοντας το τελευταίο με $(X^T X)^{-1}$ δίνει τον περιορισμένο εκτιμητή

$$\begin{aligned}\beta_r &= (X^T X)^{-1}X^T y - (X^T X)^{-1}R^T\lambda \\ &= \widehat{\beta} - (X^T X)^{-1}R^T\lambda, \qquad (3)\end{aligned}$$

όπου $\widehat{\beta} = (X^T X)^{-1}X^T y$ είναι ο μη-περιορισμένος εκτιμητής.

Πολλαπλασιάζοντας (3) με R δίνει:

$$R\beta_r = R\widehat{\beta} - R(X^T X)^{-1}R^T\lambda,$$

ή $\lambda = (R(X^TX)^{-1}R^T)^{-1}(R\widehat{\beta} - R\beta_r)$.

Θέτοντας λ στην (3) και χρησιμοποιώντας $R\beta_r = r$ δίνει τους περιορισμένους OLS εκτιμητές:

$$\boxed{\beta_r = \widehat{\beta} - (X^TX)^{-1}R^T(R(X^TX)^{-1}R^T)^{-1}(R\widehat{\beta} - r).}$$

Σημειώστε ότι έχουμε χρησιμοποιήσει $R\beta_r = r$ το οποίο υποδειλώνει ότι ο περιορισμένος εκτιμητής ελαχίστων τετραγώνων ικανοποιεί τους περιορισμούς[β΄]. Επιπλέον, ο περιορισμένος εκτιμητής β_r διαφέρει από τον μη-περιορισμένο $\widehat{\beta}$ ως προς το βαθμό στο οποίο ο τελευταίος αποτυγχάνει να ικανοποιήσει τους περιορισμούς.

Η διακύμανση των περιορισμένων OLS εκτιμητών δίνεται από:

$$\mathrm{Var}(\beta_r) = \sigma^2\Big((X^TX)^{-1} - (X^TX)^{-1}R^T\big(R(X^TX)^{-1}R^T\big)^{-1}R(X^TX)^{-1}\Big),$$

[β΄] Όντως ο παλλαπλασιασμός του τελευραίου με R δίνει $R\beta_r = r$.

ή [ϒ΄]

$$\mathrm{Var}(\beta_r) = \mathrm{Var}(\widehat{\beta}) - \sigma^2\Big((X^TX)^{-1}R^T\big(R(X^TX)^{-1}R^T\big)^{-1}R(X^TX$$

Σημειώσεις

- Ο περιορισμένος εκτιμητής ελαχίστων τετραγώνων β_r είναι αμερόληπτος MONO αν οι επιβληθέντες περιορισμοί είναι απόλυτα σωστοί.
- Η διακύμανση του περιορισμένου εκτιμητή ελαχίστων τετραγώνων β_r είναι μικρότερη από αυτή του μη-περιορισμένου OLS $\widehat{\beta}$.
- Θεωρήστε την εισαγωγή δειγματικών και εκ των προτέρων πληροφοριών στο σύστημα:

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix}\beta + \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}.$$

Ένας OLS εκτιμητής υπάρχει στη μορφή

$$\widetilde{\beta} = (X^TX + R^TR)^{-1}(X^TY + R^Tr)$$

εάν $\mathrm{rank}\begin{pmatrix} X \\ R \end{pmatrix} = n$, όπου n είναι ο αριθμός των

[ϒ΄] $\mathrm{Var}(\widehat{\beta}) = \sigma^2(X^TX)^{-1}$.

μεταβλητών.

- Οι περιορισμοί μπορεί να θεωρηθούν ως επιπλέον δεδομένα τα οποία πρέπει να ικανοποιούν με ακρίβεια τη λύση ελαχίστων τετραγώνων. Αυτό μπορεί να επιτευχθεί σταθμίζοντας τους περιορισμούς, ή αντίστοιχα σταθμίζοντας προς τα κάτω το δείγμα (δηλαδή, αφαιρώντας τους περιορισμούς) και υπολογίζοντας των OLS εκτιμητή. Ως εκ τούτου, υπολογίζουμε τους εκτιμητές ελαχίστων τετραγώνων του

$$\begin{pmatrix} \gamma y \\ r \end{pmatrix} = \begin{pmatrix} \gamma X \\ R \end{pmatrix} \beta + \begin{pmatrix} \gamma \varepsilon \\ 0 \end{pmatrix} \text{ ή, } \begin{pmatrix} y \\ \delta r \end{pmatrix} = \begin{pmatrix} X \\ \delta R \end{pmatrix} \beta + \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix},$$

όπου $\gamma \to 0$ και $\delta \to \infty$.

Παράδειγμα: Συνάρτηση παραγωγής Cobb-Douglas

```
m <- 100; s <- 1;
alpha <- 5.0; b1 <- 0.3; b2 <- 0.7;
C <- runif(m, min=100, max=400)
L <- runif(m, min=100, max=200)

Q <- alpha * C^b1 * L^b2

logC <- log(C); logL <- log(L);
e<- rnorm(m, sd=s);  logQ <- log(Q) + e;

reg <- lm(logQ ~ logC + logL);
summary(reg)

delta <- 10^4;
logQ1 <- c(logQ,delta);  logC1 <- c(logC,delta);
logL1 <- c(logL,delta);

reg1 <- lm(logQ1 ~ logC1 + logL1);
summary(reg1)

gamma <- 10^(-4);
logQ2 <- c(gamma * logQ,1);
logC2 <- c(gamma * logC,1);
logL2 <- c(gamma * logL,1);
const <- rep(gamma,101); const[m+1] <- 0;

reg2 <- lm(logQ2 ~ const + logC2 + logL2-1);
summary(reg2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3420     2.9427   0.796    0.428
logC          0.4447     0.2514   1.769    0.080
logL          0.3913     0.5003   0.782    0.436
---
Residual SE: 0.9673 on 97 DF
Multiple R-Squared:0.04, Adjusted R-squared:0.02
F-statistic: 1.764 on 2 and 97 DF, p-value:0.177

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5256     0.1693   9.009 1.71e-14
logC1         0.2534     0.2655   0.954  0.34222
logL1         0.7465     0.2654   2.812  0.00594
---
Residual SE: 1.072 on 98 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 4.3e+07 on 2 and 98 DF
p-value: < 2.2e-16

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
const   1.5248     0.1692   9.009 1.71e-14
logC2   0.2534     0.2655   0.955  0.34214
logL2   0.7466     0.2655   2.813  0.00594
---
Residual SE: 0.0001 on 98 DF
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 2.9e+07 on 3 and 98 DF
p-value: < 2.2e-16
```