Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R

Δημήτρης Φουσκάκης,
Επίκουρος Καθηγητής,
Τομέας Μαθηματικών,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Εφαρμογών,
Εθνικό Μετσόβιο Πολυτεχνείο.

Περιεχόμενα

- Εισαγωγή στη Στατιστική
- Εισαγωγή στο Στατιστικό Πακέτο R
- Περιγραφική Στατιστική
- Προσομοίωση
- Στατιστική Συμπερασματολογία
  - Ένα Δείγμα
  - Δύο Ανεξάρτητα Δείγματα
  - Δείγματα κατά Ζεύγη
  - Ποσοστά
  - Έλεγχος καλής προσαρμογής
  - Πίνακες Συνάφειας 2×2.
- Ανάλυση Παλινδρόμησης
- Ανάλυση Διασποράς

Εισαγωγή

- Αρκετές φορές σε μια στατιστική μελέτη ερχόμαστε αντιμέτωποι με το πρόβλημα της *πρόβλεψης* μιας μεταβλητής (*μεταβλητή απόκρισης*) όταν γνωρίζουμε τις τιμές κάποιας ή κάποιων άλλων μεταβλητών (*επεξηγηματικές μεταβλητές*).
- Ας θεωρήσουμε καταρχήν ότι έχουμε μία μόνο επεξηγηματική μεταβλητή.

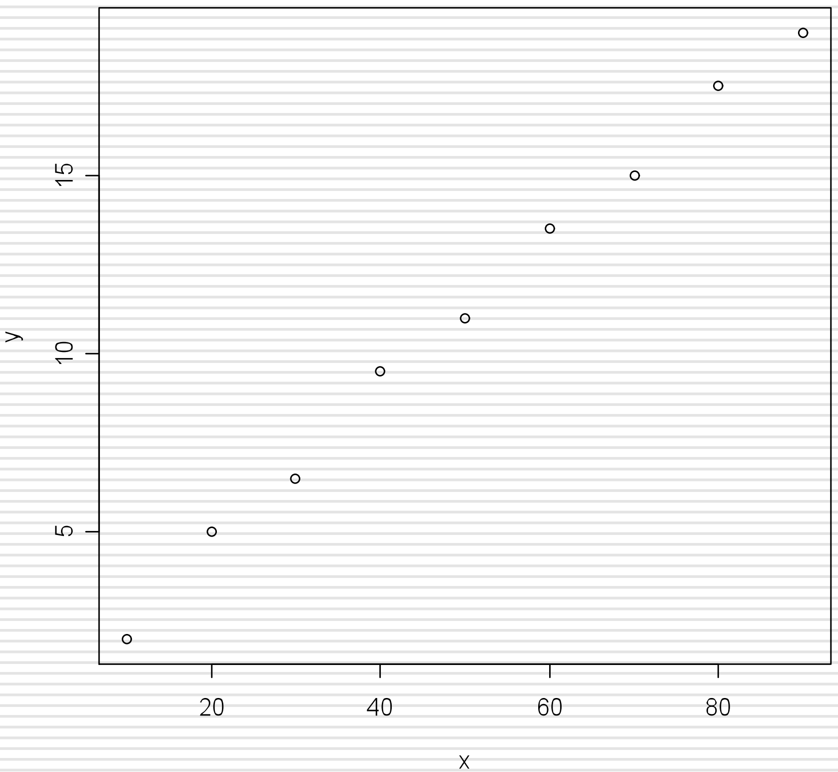
Απλό Γραμμικό Μοντέλο

- Πιο συγκεκριμένα ας υποθέσουμε ότι Υ είναι η μεταβλητή απόκρισης και Χ η επεξηγηματική μεταβλητή και ας υποθέσουμε ότι και οι δύο μεταβλητές είναι ποσοτικές. Σκοπός μας είναι να δημιουργήσουμε ένα μοντέλο, έστω Υ=g(X), έτσι ώστε στο μέλλον να μπορούμε να προσδιορίσουμε την τιμή του Υ με βάση την τιμή του Χ.

Απλό Γραμμικό Μοντέλο

- Πώς όμως επιλέγουμε την συναρτησιακή μορφή της g(X); Η επιλογή της κατάλληλης συνάρτησης g(X) μπορεί γίνει με την βοήθεια ενός τυχαίου δείγματος $(Y_1,X_1),\ldots,(Y_n,X_n)$. Πιο συγκεκριμένα αν $(y_1,x_1),\ldots,(y_n,x_n)$ οι παρατηρήσεις μας, τότε μπορούμε να σχηματίσουμε το γράφημα των σημείων (y_i,x_i), γνωστό ως διάγραμμα διασποράς των σημείων, και να εκτιμήσουμε την συναρτησιακή μορφή της g.
- Στο επόμενο γράφημα, π.χ., το διάγραμμα διασποράς υποδεικνύει ότι η σχέση της X με την Y είναι γραμμική και άρα μπορούμε να θεωρήσουμε ότι $g(X)=a+bX$. Μένει τότε να εκτιμήσουμε τις τιμές των a και b, με την βοήθεια και πάλι της πληροφορίας που έχουμε από το τυχαίο δείγμα.

Απλό Γραμμικό Μοντέλο



Απλό Γραμμικό Μοντέλο

- Η ανάλυση μας λοιπόν είναι εμπειρική (βασίζεται στην υπάρχουσα εμπειρία μας με βάση το τυχαίο δείγμα) και άρα το μοντέλο μας είναι στοχαστικό. Αντιθέτως αν η ανάλυσή μας ήταν θεωρητική, γνωρίζαμε δηλαδή όλον τον πληθυσμό, το μοντέλο θα ήταν προσδιοριστικό.
- Στα στοχαστικά μοντέλα προφανώς έχουμε ελλιπή πληροφορία, το μοντέλο στο οποίο θα καταλήξουμε μπορεί να μην ικανοποιείται ακριβώς για ζεύγη τιμών του πληθυσμού που δεν έχουν παρατηρηθεί στο τυχαίο δείγμα. Για τον λόγο αυτό στο μοντέλο προσθέτουμε και ένα τυχαίο σφάλμα ε το οποίο θεωρούμε ότι προέρχεται από μια γνωστή κατανομή με άγνωστες παραμέτρους. Το στοχαστικό δηλαδή μοντέλο παίρνει την μορφή $Y = g(X) + \varepsilon$.
- Ακούγεται λογικό να θεωρήσουμε ότι η μέση τιμή του ε είναι 0, δηλαδή κατά μέσο όρο το σφάλμα μας είναι μηδενικό. Συνήθως θεωρούμε $\varepsilon \sim N(0, \sigma^2)$, με σ^2 άγνωστο.

Απλό Γραμμικό Μοντέλο

- Ας θεωρήσουμε τώρα ότι η g είναι μια γραμμική συνάρτηση, $g(X) = a + bX$, με a, b άγνωστες ποσότητες. Δηλαδή θεωρούμε ότι η τυχαία μεταβλητή X επηρεάζει γραμμικά την αναμενόμενη τιμή της τυχαίας μεταβλητής Y.

$$Y=a+bX+\varepsilon,\ \varepsilon\sim N(0,\sigma^2) \Rightarrow E[Y \mid X]=a+bX$$

- Ισοδύναμα σε τυχαίο δείγμα $(Y_1,X_1),\ldots,(Y_n,X_n)$ θεωρούμε ότι ισχύει η σχέση

$$Y_i=a+bX_i+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma^2) \Rightarrow E[Y_i \mid X_i]=a+bX_i$$

με ε_i ανεξάρτητες (και ισόνομες) τυχαίες μεταβλητές.

- Τα ε_i καλούνται τυχαία σφάλματα, και παριστάνουν την άγνωστη κατακόρυφη απόκλιση της τιμής y_i από την ευθεία $E(Y_i \mid X_i = x_i) = a + bx_i$ για δοθείσα τιμή x_i.

Απλό Γραμμικό Μοντέλο

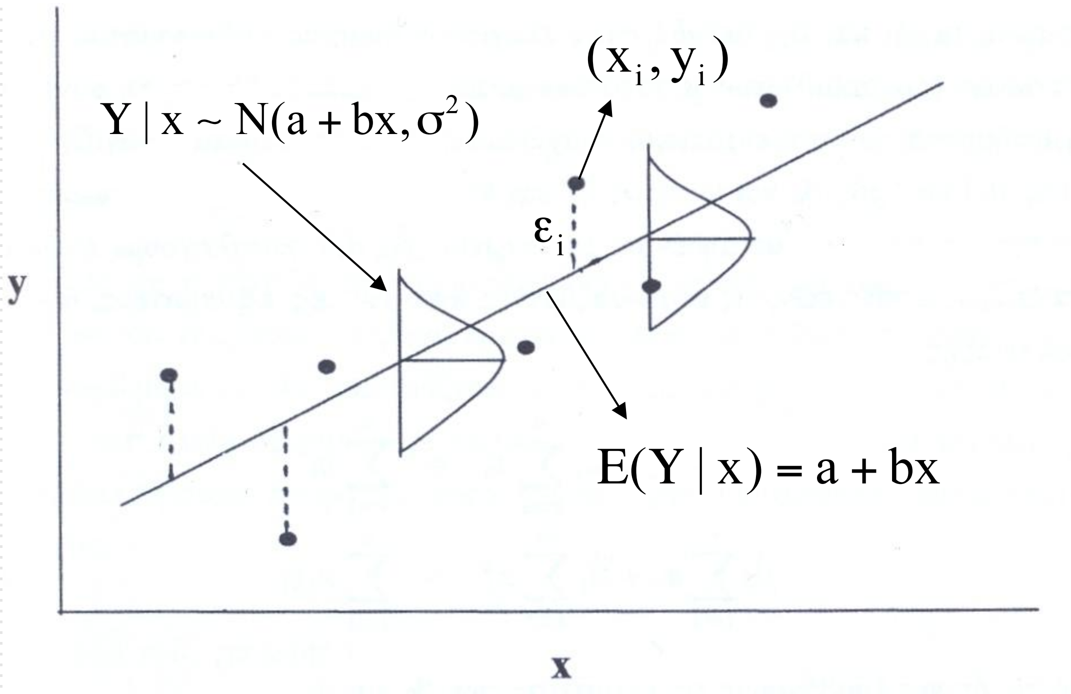
- Το παραπάνω μοντέλο καλείται απλό (γιατί έχουμε μία μόνο επεξηγηματική μεταβλητή X) γραμμικό (λόγω της γραμμικής συνάρτησης που χρησιμοποιούμε) μοντέλο παλινδρόμησης. Τα a, b και σ^2 είναι οι άγνωστες παράμετροι του μοντέλου μας (συντελεστές μοντέλου), τις οποίες θα εκτιμήσουμε με την βοήθεια των παρατηρήσεων που διαθέτουμε $(y_1,x_1),\ldots,(y_n,x_n)$ που είναι οι τιμές ενός τυχαίου δείγματος $(Y_1,X_1),\ldots,(Y_n,X_n)$.

Ερμηνεία παραμέτρων

- Η σταθερά a εκφράζει την μέση τιμή της Y όταν το X=0.
- Η σταθερά b εκφράζει το πόσο αναμένεται να μεταβληθεί η αναμενόμενη τιμή της Y, αν η X αυξηθεί κατά μία μονάδα.
- Η ποσότητα σ^2 εκφράζει την διασπορά των σφαλμάτων, την οποία θεωρούμε σταθερή ανεξάρτητα της τιμής της τ.μ. X (υπόθεση ομοσκεδαστικότητας). Επειδή η τυχαιότητα της Y δεδομένης μιας τιμής της X = x οφείλεται στα σφάλματα, το σ^2 εκφράζει και την διασπορά της δεσμευμένης κατανομής της τ.μ. $Y|x$.

Απλό Γραμμικό Μοντέλο

$$Y=a+bX+\varepsilon,\ \varepsilon \sim N(0,\sigma^2) \Leftrightarrow Y|x \sim N(a+bx,\ \sigma^2)$$



Απλό Γραμμικό Μοντέλο

- Εκτιμώντας λοιπόν τα a και b από τα $\hat{a}$ και $\hat{b}$ αντίστοιχα καταλήγουμε στο

$$\hat{Y} = \hat{a} + \hat{b}x.$$

- Το $\hat{Y}$ καλείται προβλεπόμενη τιμή και είναι όπως είδαμε η αναμενόμενη τιμή που θα πάρει η Y όταν X=x, όπως αυτήν την εκτιμήσαμε με βάση το μοντέλο παλινδρόμησης. Η προβλεπόμενη τιμή είναι τ.μ., δηλαδή για διαφορετικό δείγμα ενδέχεται να πάρει άλλη τιμή όταν X=x. Η προβλεπόμενη τιμή αποτελεί μία αμερόληπτη εκτιμήτρια της μέσης τιμής της τ.μ. Y, όταν X=x. Παρακάτω θα δούμε και ένα Δ.Ε. για την μέση τιμή αυτή όταν X=x.

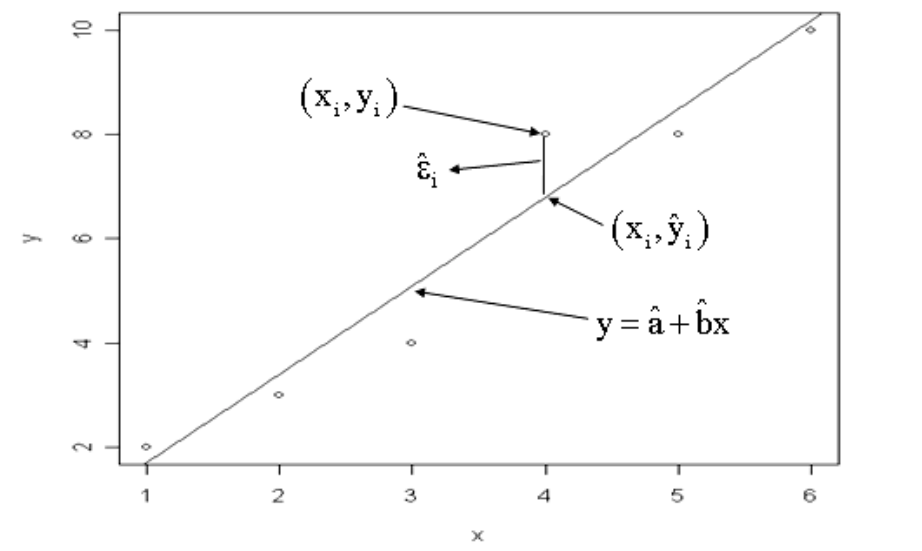
Απλό Γραμμικό Μοντέλο

- Για κάθε παρατήρηση x_i μπορούμε να υπολογίσουμε τις προβλεπόμενες τιμές
$$\hat{y}_i = \hat{a} + \hat{b}x_i.$$
- Αν η ευθεία $\hat{a} + \hat{b}x_i$ δεν περνάει ακριβώς από τα σημεία (y_i,x_i), περιμένουμε να έχουμε αποκλίσεις μεταξύ των y_i και των $\hat{y}_i$.
- Οι ποσότητες $\hat{\varepsilon}_i = y_i - \hat{y}_i$ αποτελούν τις εκτιμήσεις των άγνωστων τυχαίων σφαλμάτων ε_i και καλούνται υπόλοιπα (*residuals*).

Απλό Γραμμικό Μοντέλο

- Τους συντελεστές του μοντέλου τους εκτιμούμε με βάση τις παρατηρήσεις $(y_1,x_1),\ldots,(y_n,x_n)$, εφαρμόζοντας τη μέθοδο ελαχίστων τετραγώνων.
- Με την μέθοδο ελαχίστων τετραγώνων επιλέγουμε την ευθεία (δηλαδή τα $\hat{a}$ και $\hat{b}$) εκείνη που προσαρμόζεται καλύτερα στα δεδομένα που έχουμε. Πιο συγκεκριμένα επιλέγουμε την ευθεία εκείνη $y = \hat{a} + \hat{b}x$ που ελαχιστοποιεί το άθροισμα των $\hat{\varepsilon}_i^2$.

Απλό Γραμμικό Μοντέλο



Τα $\hat{a}$ και $\hat{b}$ τέτοια ώστε να

ελαχιστοποιείται η συνάρτηση $\sum_{i=1}^{n}\left[y_i - (a + bx_i)\right]^2 = \sum_{i=1}^{n} \varepsilon_i^2$.

Απλό Γραμμικό Μοντέλο

- Μετά από πράξεις προκύπτουν τότε οι εκτιμήτριες

$$\hat{b} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

οι οποίες είναι τυχαίες μεταβλητές (από διαφορετικό δείγμα ενδέχεται να προκύψουν διαφορετικές εκτιμήτριες).

Απλό Γραμμικό Μοντέλο

- Το σ^2 το εκτιμούμε από την ποσότητα

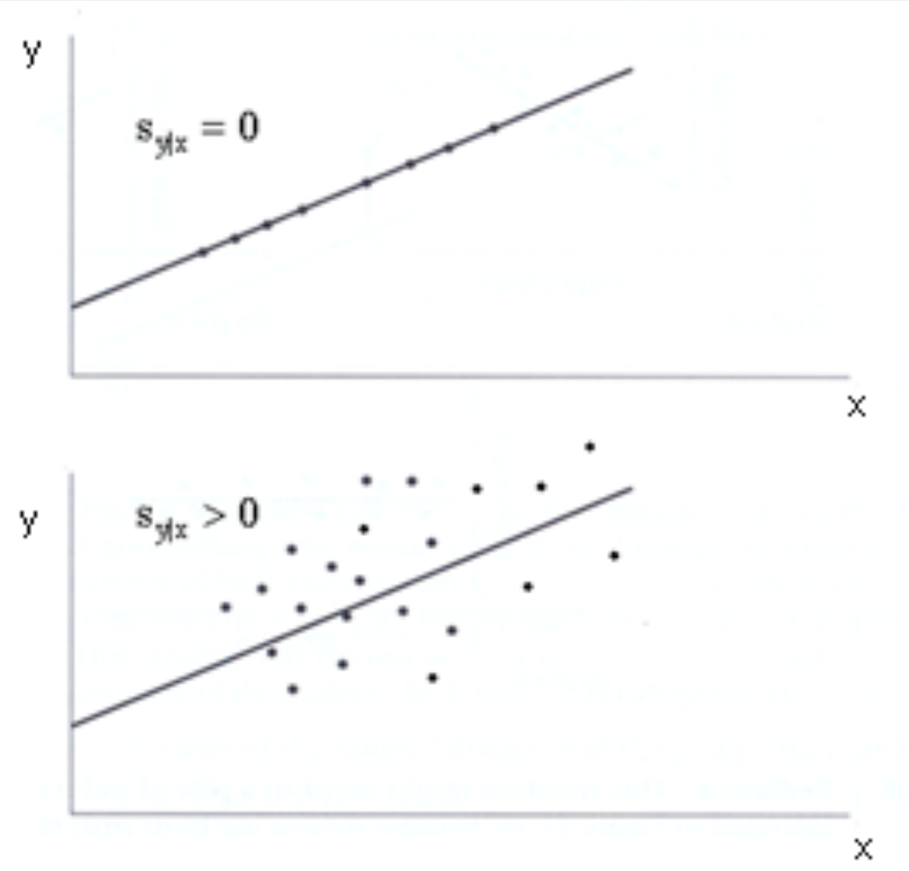
$$s_{y|x}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Εκτίμηση διασποράς των σφαλμάτων

Η θετική τετραγωνική της ρίζα της παραπάνω εκτιμήτριας καλείται τυπικό σφάλμα της παλινδρόμησης και όσο μικρότερη τιμή έχει τόσο καλύτερη προσαρμογή έχουμε για το μοντέλο παλινδρόμησης.

- Το $s_{y|x}^2$ αποτελεί αμερόληπτη εκτιμήτρια του σ^2 και καλείται μέσο τετραγωνικό σφάλμα (MSE).

Απλό Γραμμικό Μοντέλο



Συντελεστής Συσχέτισης

- Ο συντελεστής συσχέτισης (*correlation coefficient*) μεταξύ των τ.μ. X και Y εκφράζει το "βαθμό" στον οποίο μπορούμε να εκτιμήσουμε γραμμικά τη μία τ.μ. όταν γνωρίζουμε την τιμή της άλλης.

$$\rho = \mathrm{Cov}(X, Y) / \{V[X]V[Y]\}^{1/2}$$

- Όταν ρ=0 οι τ.μ. X και Y είναι ασυσχέτιστες. Όταν ρ=1 υπάρχει τέλεια θετική γραμμική συσχέτιση των δύο τ.μ. ενώ όταν ρ=-1 υπάρχει τέλεια αρνητική γραμμική συσχέτιση.

Συντελεστής Συσχέτισης

- Όταν δεν γνωρίζουμε το ρ το εκτιμούμε με την βοήθεια των παρατηρήσεων (y_i, x_i)

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\left\{\sum_1^n (x_i - x)^2 \sum_1^n (y_i - y)^2\right\}^{1/2}}$$

από το δειγματικό συντελεστή συσχέτισης.

Συντελεστής Συσχέτισης

- Ο δειγματικός συντελεστής συσχέτισης εκτιμά το βαθμό στον οποίο οι τ.μ. X και Y είναι γραμμικά συσχετισμένες, χωρίς να συνεπάγεται κατά ανάγκη κάποιο είδος αιτιακής σχέσης μεταξύ των X και Y.
- Αρκετά συχνά μας ενδιαφέρει να ελέγξουμε, σε ε.σ. έστω α%, κατά πόσο οι δύο τ.μ. X και Y είναι ασυσχέτιστες ή όχι, δηλαδή τον έλεγχο H_0: ρ=0 με εναλλακτική H_1: ρ≠0.
- Αποδεικνύεται ότι κάτω από την μηδενική υπόθεση

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim St(n-2).$$

Συντελεστής Συσχέτισης

- Υπολογίζουμε λοιπόν την τιμή του στατιστικού ελέγχου T και η P-τιμή του αμφίπλευρου ελέγχου είναι 2 φορές η πιθανότητα δεξιά του |T| με βάση την St(n-2).
- Στην R μπορούμε να εφαρμόσουμε τον εν λόγω έλεγχο με την βοήθεια της εντολής *cor.test(X,Y).*
- Αν οι δυο μεταβλητές δεν είναι κανονικές και το δείγμα δεν είναι μεγάλο τότε ο παραπάνω έλεγχος (γνωστός με το όνομα *Pearson correlation coefficient test)* δεν είναι πλέον έγκυρος. Πρέπει αντί αυτού να εφαρμοστεί ο μη παραμετρικός *Spearman rank correlation coefficient test*, κατά τον οποίο αντικαθιστούμε τις παρατηρήσεις με την σειρά κατάταξης των τιμών τους (rank) και εφαρμόζουμε την προηγούμενη μεθοδολογία.
- Στην R μπορούμε να εφαρμόσουμε τον μη παραμετρικό έλεγχο με την βοήθεια της εντολής *cor.test(X,Y, method="spearman").*
- Όταν στο δείγμα υπάρχουν ισοπαλίες (παρατηρήσεις με την ίδια τιμή) η παραπάνω εντολή μας δίνει προειδοποιητικό μήνυμα και δεν υπολογίζει την ακριβή P-τιμή του ελέγχου αλλά αυτή που προκύπτει από μια προσέγγιση.

Συντελεστής Προσδιορισμού

- Στο απλό γραμμικό μοντέλο η ποσότητα

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

καλείται συντελεστής προσδιορισμού, παίρνει τιμές στο [0,1] και εκφράζει το ποσοστό της διασποράς της τ.μ. Y που εξηγείται με βάση το μοντέλο παλινδρόμησης. Αποδεικνύεται ότι, στο απλό γραμμικό μοντέλο, η παραπάνω ποσότητα ισούται με το τετράγωνο του δειγματικού συντελεστή συσχέτισης r. Γενικά όσο μεγαλύτερες τιμές παίρνει ο συντελεστής προσδιορισμού τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των τ.μ. Y και X, υπό την προϋπόθεση ότι το γραμμικό μοντέλο είναι το κατάλληλο.

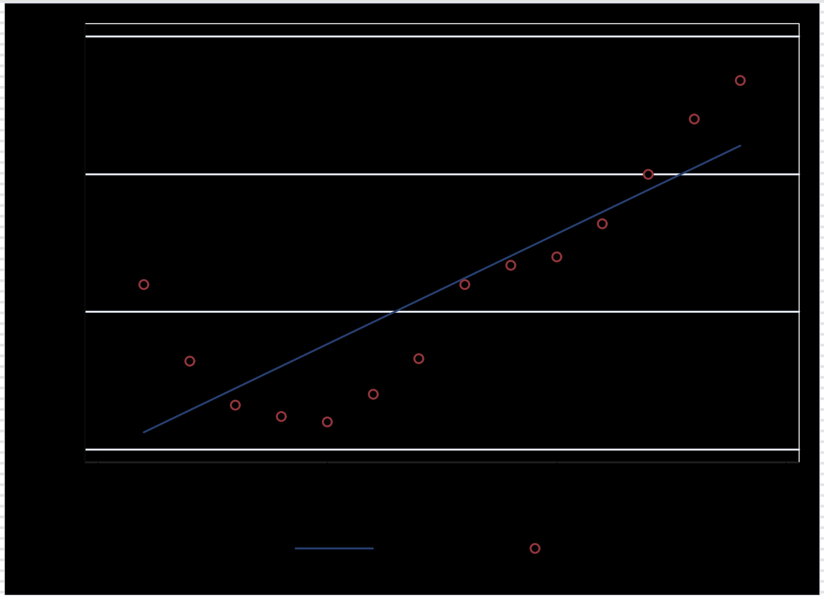
Συντελεστής Προσδιορισμού

- Αρκετές φορές υπολογίζουμε και τον διορθωμένο συντελεστή προσδιορισμού

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ n-2}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ n-1}$$

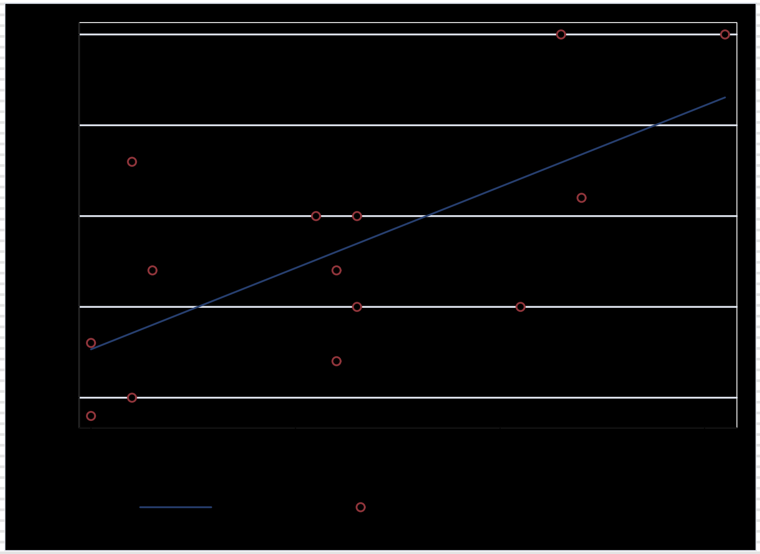
του οποίου η ερμηνεία δίνεται στο πολλαπλό γραμμικό μοντέλο.

Συντελεστής Προσδιορισμού



Εσφαλμένη υπόθεση γραμμικότητας

R^2=0.69



R^2=0.46

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Οι εκτιμήσεις των a και b που λαμβάνουμε με την μέθοδο ελαχίστων τετραγώνων βασίζονται στα συγκεκριμένα δεδομένα που διαθέτουμε. Συχνά λοιπόν ενδιαφερόμαστε να ελέγξουμε τις ακόλουθες υποθέσεις, σε ε.σ. έστω α:
  - H_0: b=0 έναντι της εναλλακτικής H_1: b $\neq$ 0
  - H_0: a=0 έναντι της εναλλακτικής H_1: a $\neq$ 0

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Με τον πρώτο έλεγχο θέλουμε να διαπιστώσουμε αν πράγματι αύξηση κατά μια μονάδα της X σημαίνει και μεταβολή της αναμενόμενης τιμής της Y.
- Στο δεύτερο έλεγχο θέλουμε να δούμε κατά πόσο η αναμενόμενη τιμή της Y είναι 0 όταν X=0. Πολλές φορές η τιμή αυτή δεν έχει ερμηνεία, διότι η τιμή X=0 δεν παρατηρείται ποτέ στην πράξη.

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Τα στατιστικά ελέγχου με βάση τις μηδενικές υποθέσεις τότε είναι:

$$T_2 = \frac{\hat{b}-0}{se(\hat{b})} = \frac{\hat{b}}{\sqrt{\sigma^2 / \sum_{i=1}^{n}(x_i-\bar{x})^2}} \approx \frac{\hat{b}}{\sqrt{s_{y|x}^2 / \sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim St(n-2)$$

$$T_1 = \frac{\hat{a}-0}{se(\hat{a})} = \frac{\hat{a}-0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \approx \frac{\hat{a}}{\sqrt{s_{y|x}{}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim St(n-2).$$

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Υπολογίζουμε λοιπόν τα παραπάνω δύο στατιστικά ελέγχου T_1 και T_2 και η P-τιμή των ελέγχων είναι 2 φορές η πιθανότητα της περιοχής της St(n -2) δεξιά από τις τιμές των στατιστικών ελέγχων που παρατηρούμε.
- Ισοδύναμα θα μπορούσαμε να είχαμε κατασκευάσει συμμετρικά (1-α)% Δ.Ε. για τα a και b και να ελέγξουμε αν η τιμή 0 ανήκει σ' αυτά τα Δ.Ε.

$$\left(\hat{b} \pm t_{n-2,\alpha/2}\, s_{y|x} \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \longrightarrow \text{(1-α)\% Δ.Ε. για το b}$$

$$\left(\hat{a} \pm t_{n-2,\alpha/2}\, s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right) \longrightarrow \text{(1-α)\% Δ.Ε. για το a}$$

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Ένας άλλος έλεγχος που συνήθως εξετάζουμε στο μοντέλο παλινδρόμησης, γνωστός με την ονομασία F-test, είναι και ο παρακάτω, ο οποίος ελέγχει κατά πόσο το προτεινόμενο μοντέλο $y=a+bx$ διαφέρει από το σταθερό $y=a$. Στη απλή γραμμική παλινδρόμηση ο εν λόγω έλεγχος είναι ισοδύναμος με τον έλεγχο για το b που είδαμε πριν.

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Υπολογίζουμε το στατιστικό ελέγχου

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

το οποίο κάτω από την μηδενική υπόθεση H_0:b=0 (με εναλλακτική H_1:b≠0) ακολουθεί την F(1,n-2). Υπολογίζουμε λοιπόν την τιμή του στατιστικού ελέγχου F και η P-τιμή είναι πιθανότητα της περιοχής της F(1,n-2) δεξιά από το F που παρατηρούμε.

Συμπερασματολογία στο Απλό Γραμμικό Μοντέλο

- Τέλος ένα συμμετρικό (1-α)% Δ.Ε. για το y όταν X=x είναι το

$$\left(\hat{y} \pm t_{n-2,\alpha/2}\, s_{y|x} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_1^n (x_i - \bar{x})^2}} \right).$$

Προϋποθέσεις απλού γραμμικού μοντέλου

1. Γραμμικότητα
2. Κανονικότητα Σφαλμάτων
3. Ομοσκεδαστικότητα
4. Ανεξαρτησία Σφαλμάτων

Παράδειγμα απλού γραμμικού μοντέλου στην R

- Μετρήσεις της ποσότητας οξειδίου Y που σχηματίζεται στην επιφάνεια ενός μετάλλου που τίθεται για χρόνο X (min) σε κλίβανο σταθερής θερμοκρασίας δίνεται από τον παρακάτω πίνακα.

x	10	20	30	40	50	60	70	80	90
y	2.0	5.0	6.5	9.5	11.0	13.5	15.0	17.5	19.0

Ζητείται να προσαρμόσουμε ένα μοντέλο απλής γραμμικής παλινδρόμησης και να ελέγξουμε κατά πόσο ο χρόνος X όπου το μέταλλο τίθεται σε κλίβανο σταθερής θερμοκρασίας επηρεάζει την ποσότητα οξειδίου Y που σχηματίζεται στην επιφάνεια του μετάλλου.

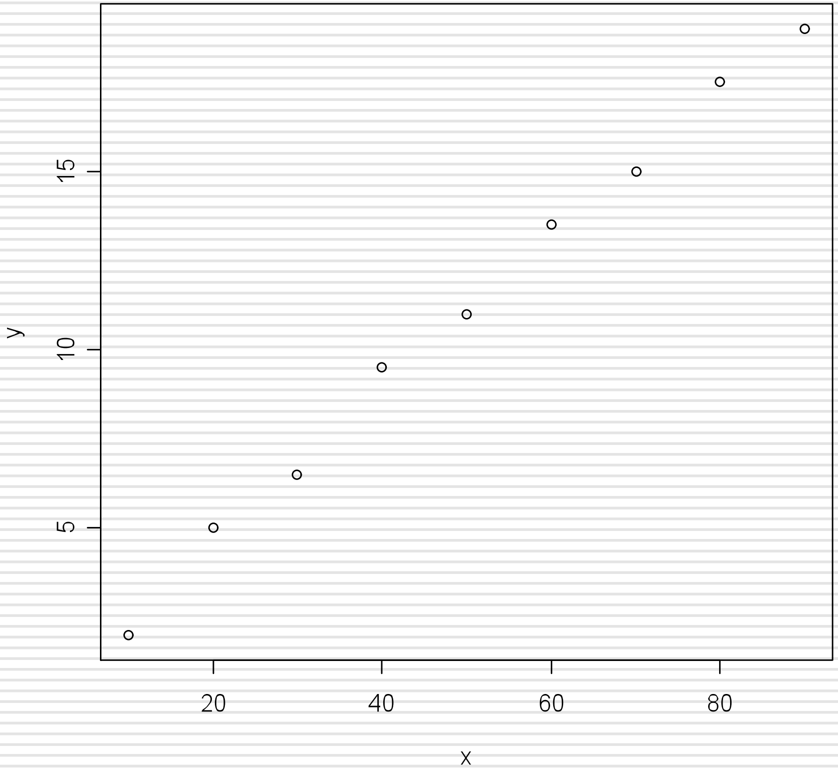
Παράδειγμα απλού γραμμικού μοντέλου στην R

```
> x<-seq(10,90,by=10)
> x
[1] 10 20 30 40 50 60 70 80 90
> y<-c(2,5,6.5,9.5,11,13.5,15,17.5,19)
```

Δημιουργούμε ένα διάγραμμα διασποράς, μια απεικόνιση δηλαδή των σημείων (x_i,y_i), για να ελέγξουμε αν η γραμμική συνάρτηση φαίνεται να είναι η κατάλληλη.

```
> plot(x,y)
```

Παράδειγμα απλού γραμμικού μοντέλου στην R



Υπόθεση γραμμικότητας λογική

Παράδειγμα απλού γραμμικού μοντέλου στην R

```
> results<-lm(y~x)
> results

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.4583       0.2108
```

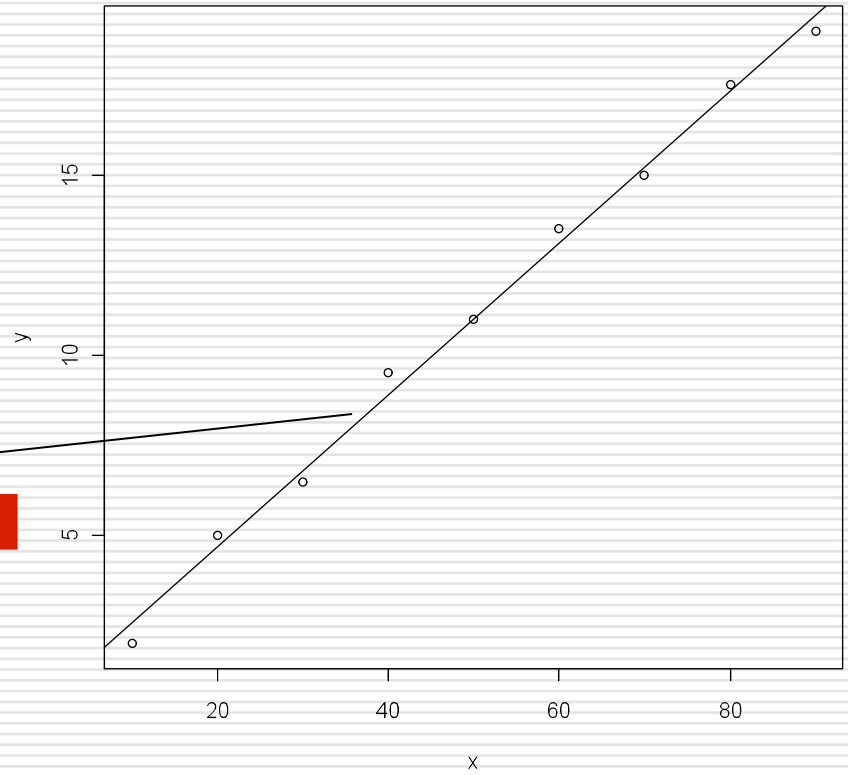
Συνάρτηση στην R που προσαρμόζει το γραμμικό μοντέλο με y τα δεδομένα για την μεταβλητή απόκρισης και x τα δεδομένα για την επεξηγηματική μεταβλητή.

$\hat{a}$

$\hat{b}$

Παράδειγμα απλού γραμμικού μοντέλου στην R

```
> plot(x,y)
> abline(results)
```



Ευθεία ελαχίστων τετραγώνων

Παράδειγμα απλού γραμμικού μοντέλου στην R

```
> summary(results)

Call:
lm(formula = y ~ x)

Residuals:
       Min         1Q     Median         3Q        Max
-5.667e-01 -2.833e-01  1.559e-16  3.250e-01  6.083e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.458333   0.312573   1.466    0.186
x           0.210833   0.005555  37.957 2.29e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4303 on 7 degrees of freedom
Multiple R-squared: 0.9952,     Adjusted R-squared: 0.9945
F-statistic:  1441 on 1 and 7 DF,  p-value: 2.292e-09
```

Περιγραφικοί δείκτες υπολοίπων

$\hat{a}$ $se(\hat{a})$ T_1

P-τιμή για τον έλεγχο του a

$\hat{b}$ $se(\hat{b})$ T_2

P-τιμή για τον έλεγχο του b

$S_{y|x}$

R^2

διορθωμένος συντελεστής προσδιορισμού

F

P-τιμή για τον F έλεγχο του b (ίδιο με το παραπάνω)

Παράδειγμα απλού γραμμικού μοντέλου στην R

```
> confint(results)
                 2.5 %    97.5 %
(Intercept) -0.2807838 1.1974504
x            0.1976989 0.2239678
```

Συμμετρικά 95% Δ.Ε. για τις παραμέτρους

Συμμετρικό 95% Δ.Ε. για το a

Συμμετρικό 95% Δ.Ε. για το b (δεν περιέχει το 0)

```
> confint(results, level=0.99)
                 0.5 %    99.5 %
(Intercept) -0.6355098 1.5521764
x            0.1913952 0.2302714
```

Συμμετρικά 99% Δ.Ε. για τις παραμέτρους

Προβλεπόμενες τιμές για κάθε x_i. Το ίδιο αποτέλεσμα προκύπτει και με την εντολή *fitted(results), όπως και με την εντολή results$fitted*.

```
> predict(results)
        1         2         3         4         5         6         7         8
 2.566667  4.675000  6.783333  8.891667 11.000000 13.108333 15.216667 17.325000
        9
19.433333
```

Υπόλοιπα. Το ίδιο αποτέλεσμα προκύπτει και με την εντολή *results$res*.

```
> residuals(results)
            1             2             3             4             5
-5.666667e-01  3.250000e-01 -2.833333e-01  6.083333e-01  1.558947e-16
            6             7             8             9
 3.916667e-01 -2.166667e-01  1.750000e-01 -4.333333e-01
```

Παράδειγμα απλού γραμμικού μοντέλου στην R

```
> predict(results, int="c")
        fit       lwr       upr
1  2.566667  1.941342  3.191992
2  4.675000  4.155123  5.194877
3  6.783333  6.354364  7.212303
4  8.891667  8.527990  9.255343
5 11.000000 10.660870 11.339130
6 13.108333 12.744657 13.472010
7 15.216667 14.787697 15.645636
8 17.325000 16.805123 17.844877
9 19.433333 18.808008 20.058658
```

Προβλεπόμενες τιμές των y και συμμετρικά 95% Δ.Ε. για κάθε x_i

```
> predict(results,list(x=c(15,47)), int="c")
        fit       lwr       upr
1  3.620833  3.049573  4.192094
2 10.367500 10.026088 10.708912
```

Προβλεπόμενες τιμές των y και συμμετρικά 95% Δ.Ε. όταν x=15 και x=47

```
> predict(results,list(x=c(15,47)), int="c", level=0.99)
        fit      lwr      upr
1  3.620833 2.775406  4.46626
2 10.367500 9.862234 10.87277
```

Προβλεπόμενες τιμές των y και συμμετρικά 99% Δ.Ε. όταν x=15 και x=47

Παράδειγμα απλού γραμμικού μοντέλου στην R

- Παρατηρούμε λοιπόν ότι

$$\hat{y} = 0.46 + 0.21x.$$

- Ακόμα παρατηρούμε ότι ο συντελεστής προσδιορισμού R^2= 0.99, έχουμε δηλαδή σχεδόν τέλεια προσαρμογή του μοντέλου, ενώ $S_{y|x}$=0.42 (αρκετά μικρό.)

- Τέλος θα μπορούσαμε να είχαμε υπολογίσει τον δειγματικό συντελεστή συσχέτισης με την βοήθεια της εντολής cor(x,y)

```
> cor(x,y)
[1] 0.9975795
> cor(x,y)^2
[1] 0.9951648
```

$=R^2$

Παράδειγμα απλού γραμμικού μοντέλου στην R

- Στον έλεγχο για το b βλέπουμε ότι η P-τιμή είναι πολύ μικρή οπότε πράγματι αύξηση κατά μια μονάδα της X σημαίνει και μεταβολή της αναμενόμενης τιμής της Y. Πιο συγκεκριμένα αύξηση του χρόνου κατά 1 λεπτό σημαίνει αύξηση της αναμενόμενης ποσότητας οξειδίου κατά 0.21.
- Αντιθέτως η σταθερά δεν φαίνεται από την p-τιμή του αντίστοιχου ελέγχου να είναι στατιστικά διάφορη του μηδενός.
- Για να ελέγξουμε τις προϋποθέσεις του μοντέλου κάνουμε τα εξής:

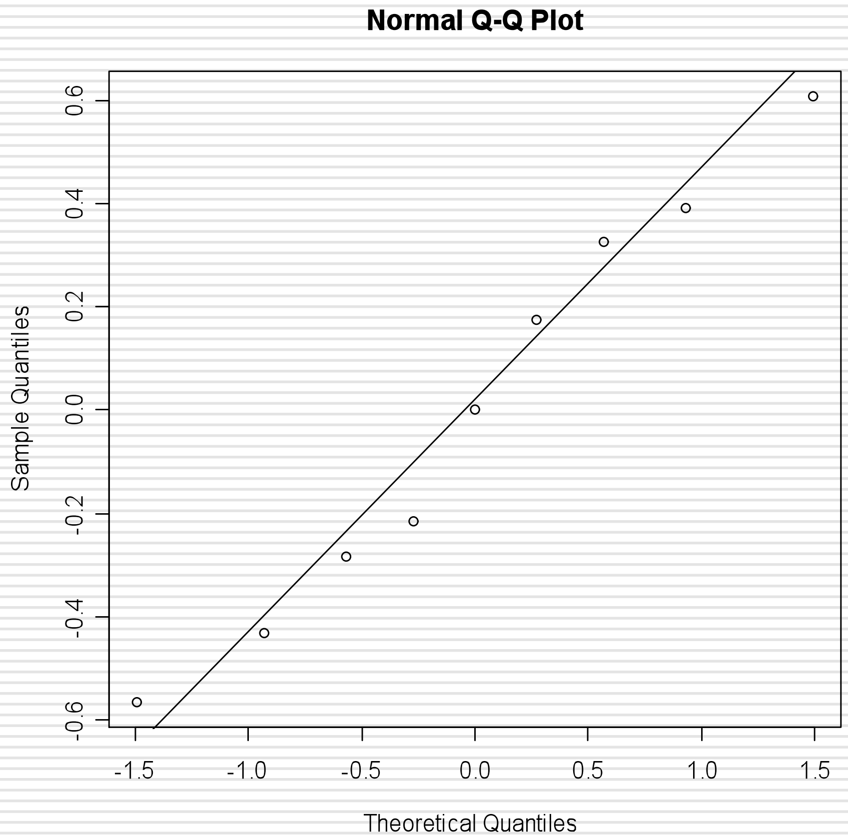
Παράδειγμα απλού γραμμικού μοντέλου στην R

- Γραμμικότητα. Ήδη ελέγχθηκε με το διάγραμμα διασποράς.
- Κανονικότητα σφαλμάτων. Ιστογράμματα και QQplots για τα υπόλοιπα. Π.χ.

```
> qqnorm(results$res)
> qqline(results$res)
```

υπόλοιπα

Παράδειγμα απλού γραμμικού μοντέλου στην R



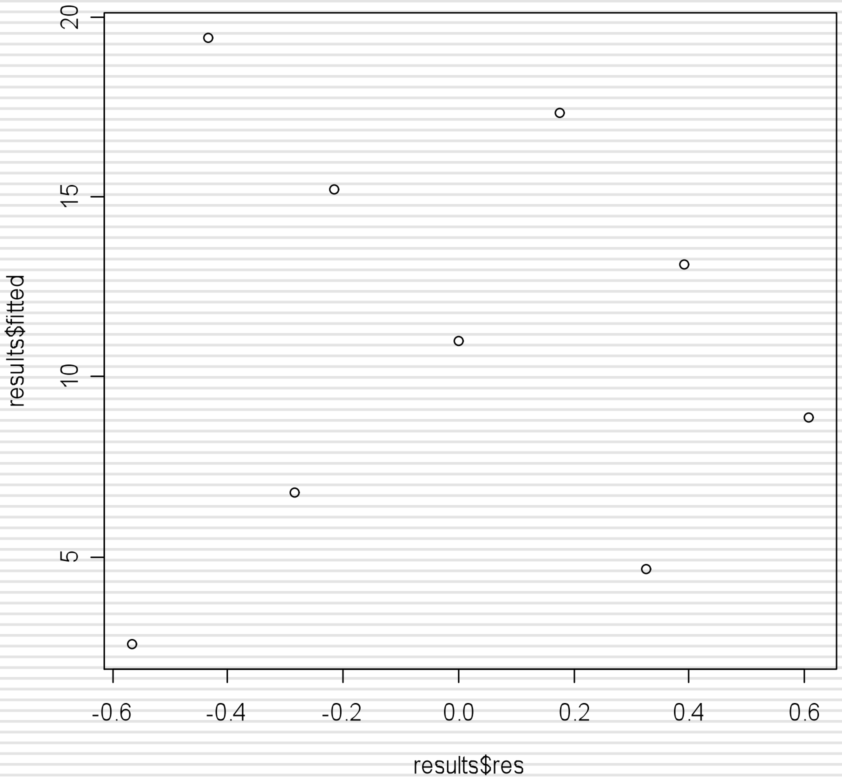
Παράδειγμα απλού γραμμικού μοντέλου στην R

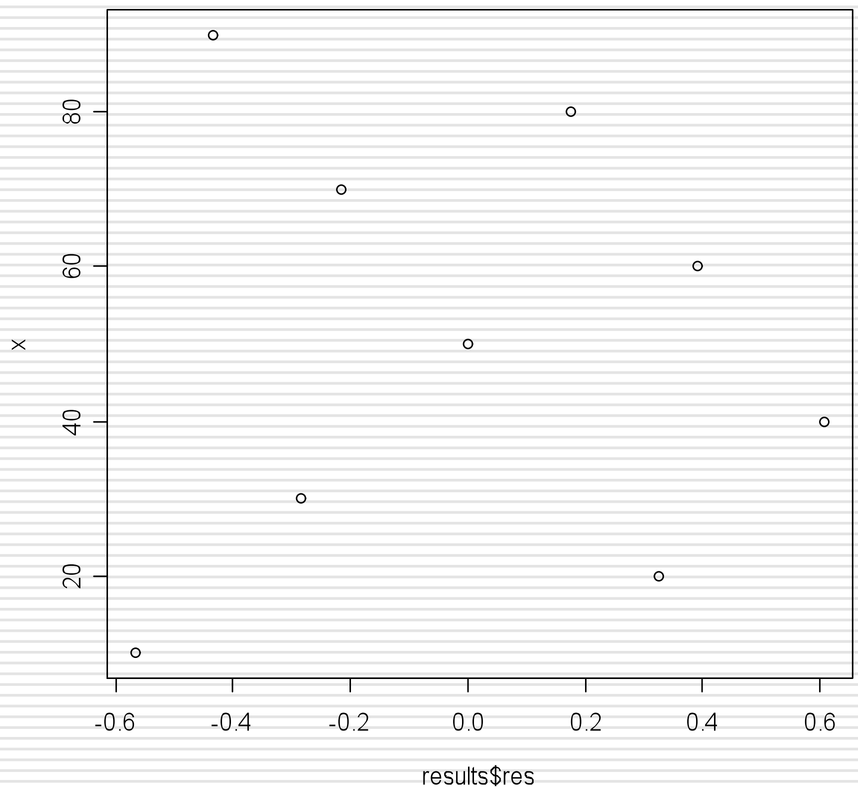
- Ομοσκεδαστικότητα. Γραφική παράσταση των υπολοίπων συναρτήσει των προβλεπόμενων τιμών ή συναρτήσει των τιμών της X. Τα ζεύγη αυτών των τιμών δεν πρέπει να εμφανίζουν κάποιο συστηματικό τρόπο συμπεριφοράς.

```
> plot(results$res, results$fitted)
> plot(results$res, x)
```

προβλεπόμενες τιμές

Παράδειγμα απλού γραμμικού μοντέλου στην R



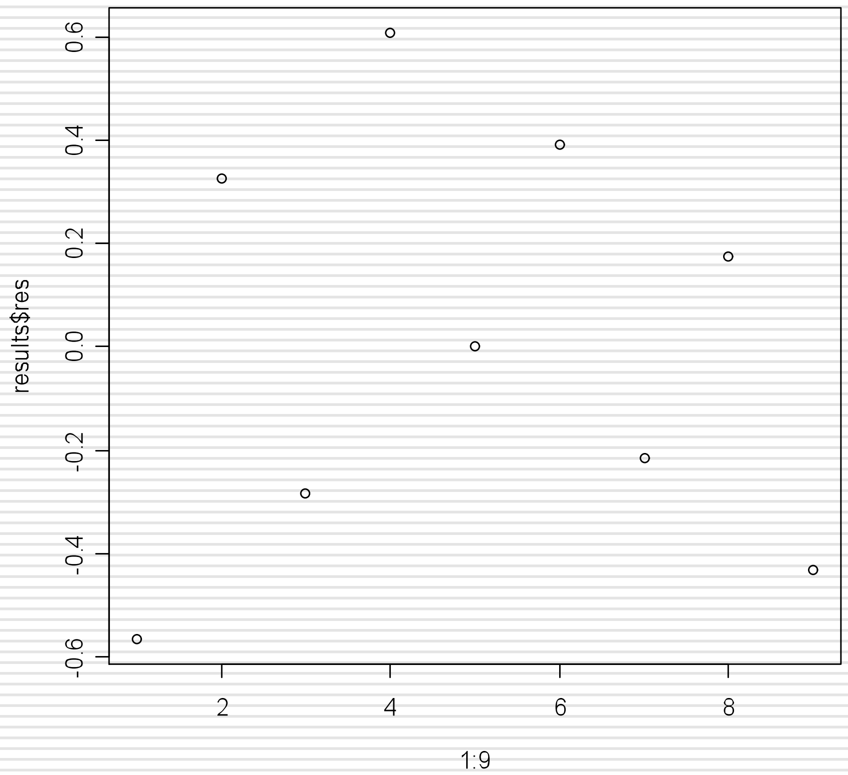


Παράδειγμα απλού γραμμικού μοντέλου στην R

- Ανεξαρτησία σφαλμάτων. Κατασκευάζουμε ένα διάγραμμα υπολοίπων σε σχέση με την σειρά των δεδομένων, στο οποίο δεν πρέπει να παρουσιάζεται κάποια σχέση και τα υπόλοιπα να συμπεριφέρονται τυχαία.

```
> plot(1:9,results$res)
```

Παράδειγμα απλού γραμμικού μοντέλου στην R



Παράδειγμα απλού γραμμικού μοντέλου στην R

- Διάφορα διαγνωστικά διαγράμματα μπορούν να γίνουν στην R με την βοήθεια της εντολής

```
> plot(results)
```

Γενικό Γραμμικό Μοντέλο

- Ας θεωρήσουμε τώρα ότι διαθέτουμε p επεξηγηματικές μεταβλητές **X**=(X_1,...,X_p), έστω όλες ποσοτικές, οι οποίες συνδέονται πάλι γραμμικά με την ποσοτική μεταβλητή απόκρισης Y.

$$Y = a + b_1 X_1 + ... + b_p X_p + \varepsilon,\ \varepsilon \sim N(0, \sigma^2) \Leftrightarrow$$

$$E(Y \mid X_1, ..., X_p) = a + b_1 X_1 + ... + b_p X_p$$

Γενικό ή πολλαπλό γραμμικό μοντέλο παλινδρόμησης

Γενικό Γραμμικό Μοντέλο

- Έστω $(Y_1,X_{11},...,X_{p1}),...,(Y_n,X_{1n},...,X_{pn})$ τυχαίο δείγμα. Τότε ισοδύναμα:

$$Y_i = a + b_1 X_{1i} + ... + b_p X_{pi} + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2) \Leftrightarrow$$

$$E(Y_i \mid X_{1i},...,X_{pi}) = a + b_1 X_{1i} + ... + b_p X_{pi}$$

όπου ε_i ανεξάρτητα.

- Με την βοήθεια των παρατηρήσεων $(y_1,x_{11},...,x_{p1}),...,(y_n,x_{1n},...,x_{pn})$ και με βάση την μέθοδο ελαχίστων τετραγώνων παίρνουμε πάλι εκτιμήτριες για τις παραμέτρους του μοντέλου a, $b_1,...,b_p$, σ^2.

Ερμηνεία Παραμέτρων

- Η παράμετρος a εκφράζει την μέση τιμή της τ.μ. Y όταν όλα τα X_j, j=1,…,p, είναι μηδέν.
- Η παράμετρος b_j, j=1,…,p, εκφράζει την αναμενόμενη μεταβολή της τιμή της τ.μ. Y όταν η X_j αυξηθεί κατά μία μονάδα και οι υπόλοιπες X_k, k≠j, παραμείνουν σταθερές.
- Η ποσότητα σ^2 εκφράζει την διασπορά των σφαλμάτων, την οποία θεωρούμε σταθερή ανεξάρτητα των τιμών των τ.μ. $X_1,\dots,X_p$ (υπόθεση ομοσκεδαστικότητας). Επειδή η τυχαιότητα της Y δεδομένου των τιμών των X οφείλεται στα σφάλματα, το σ^2 εκφράζει και την διασπορά της δεσμευμένης κατανομής της τ.μ. $Y|\mathbf{X}$.

Γενικό Γραμμικό Μοντέλο με πίνακες

$$\mathbf{Y}=\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix},\quad \tilde{\mathbf{X}}=\begin{pmatrix}1 & X_{11} & \cdots & X_{1p}\\ \vdots & X_{21} & \ddots & \vdots\\ \vdots & \vdots & & \\ 1 & X_{n1} & \cdots & X_{np}\end{pmatrix},\quad \mathbf{b}=\begin{pmatrix}a\\b_1\\\vdots\\b_p\end{pmatrix},\quad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix}$$

$$\mathbf{Y}=\tilde{\mathbf{X}}\mathbf{b}+\boldsymbol{\varepsilon}$$

- Η εκτιμήτριες με την μέθοδο ελαχίστων τετραγώνων είναι:

$$\hat{b}=\left(\tilde{X}^{T}\tilde{X}\right)^{-1}\tilde{X}^{T}Y \quad \text{και} \quad s^2_{y|x_1,\ldots,x_p}=\frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

οι οποίες είναι τυχαίες μεταβλητές (από διαφορετικό δείγμα ενδέχεται να προκύψουν διαφορετικές εκτιμήτριες).

Γενικό Γραμμικό Μοντέλο

- Εκτιμώντας λοιπόν το **b** από το $\hat{b}$ καταλήγουμε στο

$$\hat{Y} = \tilde{\mathbf{X}}\hat{\mathbf{b}}.$$

- Το $\hat{Y}$ καλείται προβλεπόμενη τιμή και είναι η αναμενόμενη που θα πάρει η τ.μ. Υ όταν **X**=**x**, όπως αυτήν την εκτιμήσαμε με βάση το μοντέλο παλινδρόμησης. Η προβλεπόμενη τιμή είναι τ.μ., δηλαδή για διαφορετικό δείγμα ενδέχεται να πάρει άλλη τιμή όταν **X**=**x**. Η προβλεπόμενη τιμή αποτελεί μία αμερόληπτη εκτιμήτρια της άγνωστης τιμής y όπου παίρνει η τ.μ. Υ όταν **X**=**x**. Παρακάτω θα δούμε ένα Δ.Ε. για την τιμή αυτή y όταν **X**=**x**.

Γενικό Γραμμικό Μοντέλο

- Για κάθε $x_{1i},\ldots,x_{pi}$ μπορούμε να υπολογίσουμε τις προβλεπόμενες τιμές
$$\hat{y}_i = \hat{a} + \hat{b}_1 x_{1i} + \ldots + \hat{b}_p x_{pi}.$$
- Οι ποσότητες $\hat{\varepsilon}_i = y_i - \hat{y}_i$ αποτελούν τις εκτιμήσεις των σφαλμάτων των μετρήσεων και καλούνται όπως και πριν υπόλοιπα.

Συντελεστής πολλαπλού προσδιορισμού

- Στο γενικό γραμμικό μοντέλο η ποσότητα

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

  καλείται πολλαπλός συντελεστής προσδιορισμού, παίρνει τιμές στο [0,1] και εκφράζει το ποσοστό της διασποράς της τ.μ. Y που εξηγείται με βάση το μοντέλο παλινδρόμησης.
- Κάθε φορά που προσθέτουμε μια επεξηγηματική μεταβλητή στο μοντέλο ο πολλαπλός συντελεστής προσδιορισμού αυξάνεται. Όταν δεν θέλουμε να υπάρχει αυτή η εξάρτηση από το p υπολογίζουμε τον διορθωμένο συντελεστής προσδιορισμού

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ n-p-1}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ n-1}$$

Συμπερασματολογία στο γενικό γραμμικό μοντέλο

- Ενδιαφερόμαστε όπως και πριν για τους εξής ελέγχους υποθέσεων:
  - H_0: a=0 έναντι της εναλλακτικής H_1: $a \neq 0$
  - H_0: b_1=0 έναντι της εναλλακτικής H_1: $b_1 \neq 0$

    . . . .

    . . . .

    . . . .
  - H_0: b_p=0 έναντι της εναλλακτικής H_1: $b_p \neq 0$
- Επίσης ενδιαφέρον έχει και το F-test το οποίο ελέγχει την μηδενική υπόθεση H_0: $b_1 = b_2 = \ldots = b_p = 0$, με εναλλακτική ότι τουλάχιστον ένα από τα b_j δεν είναι 0.

Προϋποθέσεις γενικού γραμμικού μοντέλου

1. Γραμμικότητα
2. Κανονικότητα Σφαλμάτων
3. Ομοσκεδαστικότητα
4. Ανεξαρτησία Σφαλμάτων

Προϋποθέσεις γενικού γραμμικού μοντέλου

- Γραμμικότητα. Στην απλή γραμμική παλινδρόμηση ο έλεγχος της γραμμικότητας γινόταν με το διάγραμμα διασποράς των (y_i, x_i). Στην πολλαπλή παλινδρόμηση, με p επεξηγηματικές μεταβλητές, θα μπορούσαμε να δημιουργήσουμε p διαφορετικά διαγράμματα διασποράς, ένα για κάθε επεξηγηματική μεταβλητή, και να εξετάσουμε την υπόθεση της γραμμικότητας. Βέβαια με τον τρόπο αυτόν δεν ελέγχουμε την εγκυρότητα του γενικού γραμμικού μοντέλου αλλά ελέγχουμε την εγκυρότητα των παρακάτω απλών γραμμικών μοντέλων:

$$E[Y|x_j] = a + b_j x_j,\ \ j = 1, \ldots, p.$$

Προϋποθέσεις γενικού γραμμικού μοντέλου

- Αν οι επεξηγηματικές μεταβλητές είναι ασυσχέτιστες, τότε δεν υπάρχει διαφορά από το να ελέγξουμε τα παραπάνω μοντέλα ή το γενικό γραμμικό μοντέλο. Αλλά στις περισσότερες περιπτώσεις οι επεξηγηματικές μεταβλητές συσχετίζονται. Τότε αυτό που πρέπει να ελέγξουμε είναι αν πράγματι η επεξηγηματική μεταβλητή X_j, $j = 1,\ldots,p$, συνδέεται γραμμικά με την δεσμευμένη μέση τιμή της Υ, αν όλες οι άλλες επεξηγηματικές τ.μ. $X_1,\ldots,X_{j-1}$, $X_{j+1},\ldots,X_p$ συνδέονται γραμμικά με την δεσμευμένη μέση τιμή της Υ. Με άλλα λόγια αν θεωρήσουμε την σχέση

$$y_i \approx \hat{a} + \hat{b}_1 x_{i1} + \ldots + \hat{b}_{j-1} x_{i,j-1} + p_j(x_{ij}) + \hat{b}_{j+1} x_{i,j+1} + \ldots + \hat{b}_p x_{ip} \quad (i = 1,\ldots,n),$$

χρησιμοποιώντας τις εκτιμήτριες ελαχίστων τετραγώνων που έχουμε από το γενικό γραμμικό μοντέλο, αρκεί να δείξουμε ότι η συνάρτηση p_j είναι γραμμική.

Προϋποθέσεις γενικού γραμμικού μοντέλου

- Αν τώρα θεωρήσουμε την σχέση που μας δίνει τα υπόλοιπα

$$y_i = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots + \hat{b}_p x_p + \hat{\varepsilon}_i \quad (i = 1, \ldots, n)$$

και την αφαιρέσουμε από την σχέση στην προηγούμενη διαφάνεια έχουμε

$$p_j(x_{ij}) \approx \hat{b}_j x_{ij} + \hat{\varepsilon}_i \equiv P_{ij} \quad (i = 1, \ldots, n).$$

Προϋποθέσεις γενικού γραμμικού μοντέλου

- Οι όροι P_{ij} καλούνται j–μερικά υπόλοιπα (*partial residuals*). Από την τελευταία σχέση είναι εμφανές ότι μπορούμε να ελέγξουμε αν η συνάρτηση p_j είναι γραμμική, με το διάγραμμα διασποράς των σημείων (x_{ij}, P_{ij}), $i=1,\ldots,n$ Επαναλαμβάνοντας την διαδικασία για κάθε $j=1,\ldots,p$ ελέγχουμε την γραμμικότητα στο γενικό γραμμικό μοντέλο.
- Οι υπόλοιπες προϋποθέσεις ελέγχονται όπως και στο απλό γραμμικό μοντέλο.

Παράδειγμα γενικού γραμμικού μοντέλου στην R

- Η αντοχή Y ξύλινων δοκών, σε σχέση με την περιεκτικότητα τους σε νερό X_1 και το ειδικό τους βάρος X_2, αποτέλεσε το αντικείμενο μιας μελέτης. Με βάση τα παρακάτω δεδομένα, να προσαρμοστεί το μοντέλο $Y=a+b_1X_1+b_2X_2$ και να ερμηνευτούν τα αποτελέσματά του.

Παράδειγμα γενικού γραμμικού μοντέλου στην R

Δοκός	Αντοχή Y	Περιεκτικότητα σε Νερό X_1	Ειδικό Βάρος X_2
1	11.14	11.1	0.499
2	12.74	8.9	0.558
3	13.13	8.8	0.604
4	11.51	8.9	0.441
5	12.38	8.8	0.550
6	12.60	9.9	0.528
7	11.13	10.7	0.418
8	11.70	10.5	0.480
9	11.02	10.5	0.406
10	11.41	10.7	0.467

Παράδειγμα γενικού γραμμικού μοντέλου στην R

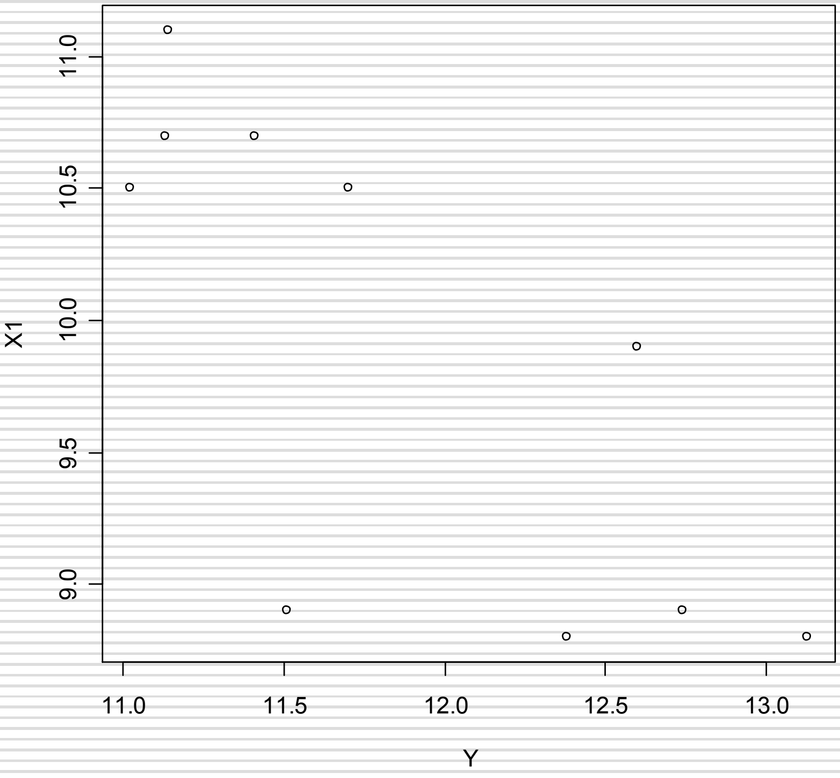
```
> Y<-c(11.14, 12.74, 13.13, 11.51, 12.38,
   12.60, 11.13, 11.70, 11.02, 11.41)
> X1<-c(11.1, 8.9, 8.8, 8.9, 8.8, 9.9, 10.7,
   10.5, 10.5, 10.7)
> X2<-c(0.499, 0.558, 0.604, 0.441, 0.55,
   0.528, 0.418, 0.48, 0.406, 0.467)
> cor(X1,X2)
[1] -0.6077351
```

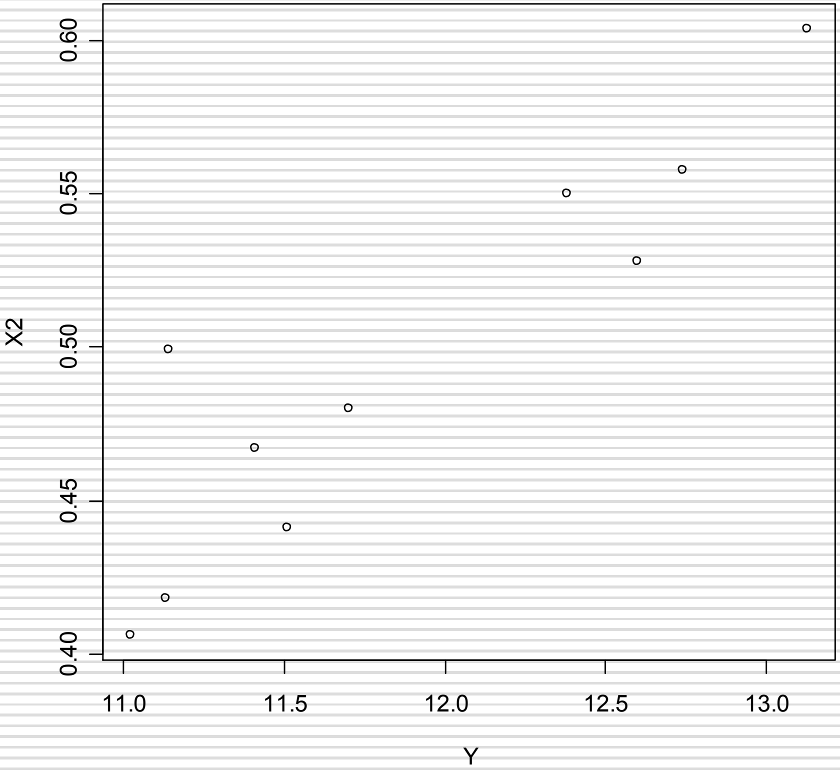
παρατηρούμε ότι X_1 και X_2 συσχετίζονται

Με τα επόμενα γραφήματα δεν μπορούμε να ελέγξουμε υπόθεση γραμμικότητας

Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> plot(Y,X1)
> plot(Y,X2)
```





Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> results<-lm(Y~X1+X2)
> results
Call:
lm(formula = Y ~ X1 + X2)
Coefficients:
(Intercept)          X1          X2
   10.3015      -0.2663      8.4947
> residuals(results)

  -0.44421731  0.06868776  0.04129893 -0.16743109
 -0.24998669  0.44985044  0.12732572  0.11738938
 0.06599797 -0.00891511
```

Προσαρμόζουμε το μοντέλο $Y=a+b_1X_1+b_2X_2$

$\hat{a}$ $\hat{b}_1$ $\hat{b}_2$

Ισοδύναμο με την εντολή results$res

Υπόλοιπα

Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> residuals(results, "partial")
           X1          X2
1  -0.76912937 -0.41108794
2   0.32968269  0.60300506
3   0.32892600  0.96637293
4   0.09356385 -0.62699494
5   0.03764038  0.21637293
6   0.44452402  0.72932643
7  -0.09105780 -0.52761648
8  -0.04772987 -0.01088075
9  -0.09912127 -0.69088075
10 -0.22729863 -0.24761648
attr(,"constant")
[1] 11.876
```

Μερικά υπόλοιπα. Για να τα υπολογίσει η R μετασχηματίζει πρώτα τις X_1 και X_2 αφαιρώντας τον μέσο τους και διαιρώντας με την τυπική τους απόκλιση

```
> newx1<-(X1-mean(X1))/sd(X1)
> newx2<-(X2-mean(X2))/sd(X2)
> lm(Y~newx1+newx2)
Call:
lm(formula = Y ~ newx1 + newx2)

Coefficients:
(Intercept)       newx1       newx2
    11.8760     -0.2488      0.5502
```

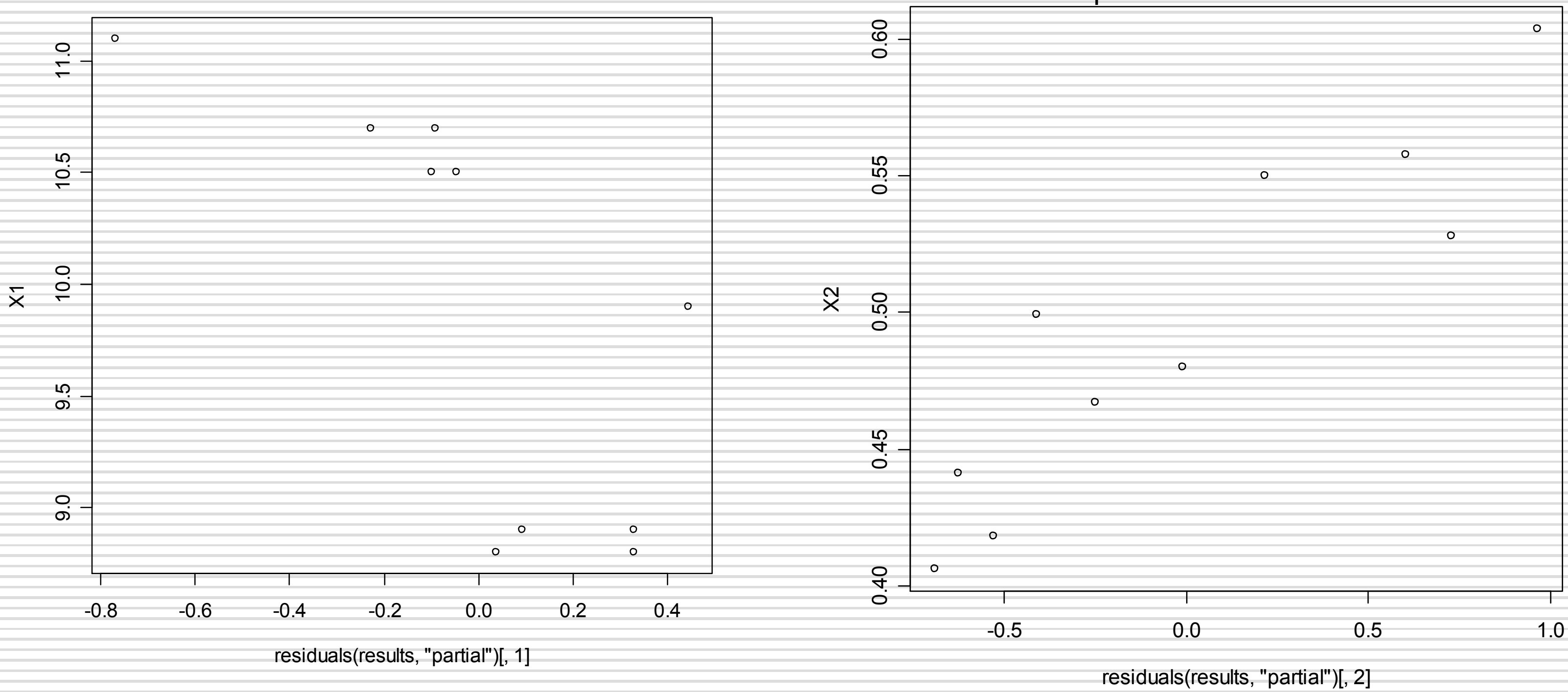
```
> -0.44421731-0.2488*newx1[1]
[1] -0.7691031
```

$\hat{\varepsilon}_i + \hat{b}_1 x_{11}$

Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> plot(residuals(results, "partial")[,1],X1)
> plot(residuals(results, "partial")[,2],X2)
```

υπόθεση γραμμικότητας λογική



Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> summary(results)
Call:
lm(formula = Y ~ X1 + X2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.44422 -0.12780  0.05365  0.10521  0.44985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3015     1.8965   5.432 0.000975 ***
X1           -0.2663     0.1237  -2.152 0.068394 .
X2            8.4947     1.7850   4.759 0.002062 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2754 on 7 degrees of freedom
Multiple R-squared:   0.9,     Adjusted R-squared: 0.8714
F-statistic:  31.5 on 2 and 7 DF,  p-value: 0.0003163
```

Περιγραφικοί δείκτες υπολοίπων

$\hat{a}$ $se(\hat{a})$

P-τιμή για τον έλεγχο του a

$\hat{b}_1$ P-τιμή για τον έλεγχο του b_1

$\hat{b}_2$ P-τιμή για τον έλεγχο του b_2

$se(\hat{b}_1)$ $se(\hat{b}_1)$

$S_{y|x}$

R^2

διορθωμένος συντελεστής προσδιορισμού

F

P-τιμή για τον F έλεγχο

Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> confint(results)
                 2.5 %     97.5 %
(Intercept)  5.8170162 14.7860314
X1          -0.5589333  0.0262906
X2           4.2738001 12.7156214
```

Συμμετρικά 95% Δ.Ε. για τις παραμέτρους

Συμμετρικό 95% Δ.Ε. για το a

Συμμετρικό 95% Δ.Ε. για το b_1 (περιέχει το 0)

Συμμετρικό 95% Δ.Ε. για το b_2 (δεν περιέχει το 0)

```
> predict(results)
       1        2        3        4        5        6        7        8
11.58422 12.67131 13.08870 11.67743 12.62999 12.15015 11.00267 11.58261
       9       10
10.95400 11.41892
```

Προβλεπόμενες τιμές για κάθε $\mathbf{x}_i$

```
> residuals(results)
          1           2           3           4           5           6
-0.44421731  0.06868776  0.04129893 -0.16743109 -0.24998669  0.44985044
          7           8           9          10
 0.12732572  0.11738938  0.06599797 -0.00891511
```

Υπόλοιπα

Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> predict(results, int="c")
       fit      lwr      upr
1  11.58422 11.16318 12.00526
2  12.67131 12.35100 12.99163
3  13.08870 12.66797 13.50943
4  11.67743 11.17108 12.18378
5  12.62999 12.30291 12.95706
6  12.15015 11.89970 12.40060
7  11.00267 10.66952 11.33582
8  11.58261 11.32699 11.83823
9  10.95400 10.58816 11.31985
10 11.41892 11.13701 11.70082
```

Προβλεπόμενες τιμές των y και συμμετρικά 95% Δ.Ε. για κάθε $\mathbf{x}_i$

```
> predict(results,list(X1=c(10,11), X2=c(0.42, 0.48)), int="c")
       fit      lwr      upr
1 11.20609 10.84469 11.56749
2 11.44945 11.09098 11.80792
```

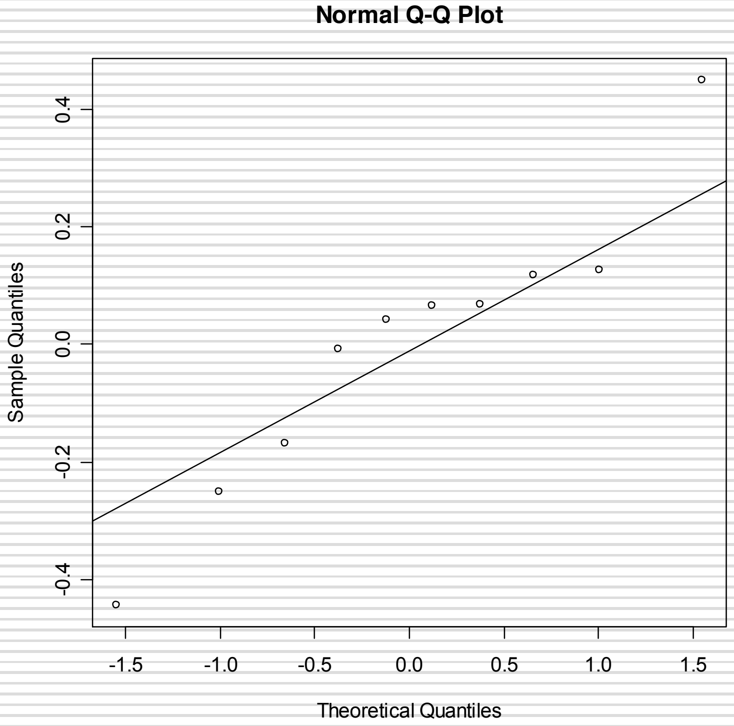
Προβλεπόμενες τιμές των y και συμμετρικά 95% Δ.Ε. όταν x_1=10 και x_2 =0.42 και όταν x1=11 και x2 =0.48

Παράδειγμα γενικού γραμμικού μοντέλου στην R

Προϋποθέσεις

```
> qqnorm(residuals(results))
> qqline(residuals(results))
```

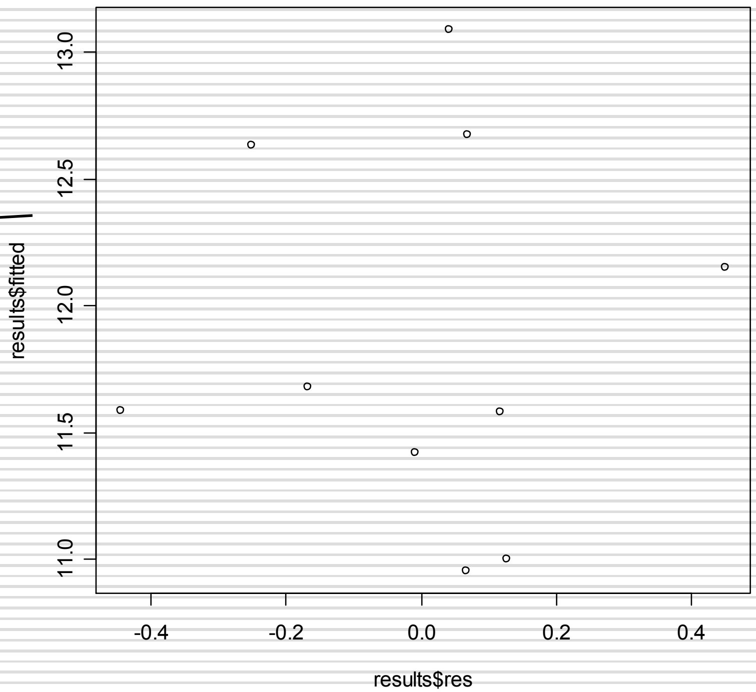
κανονικότητα υπολοίπων



Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> plot(results$res, results$fitted)
```

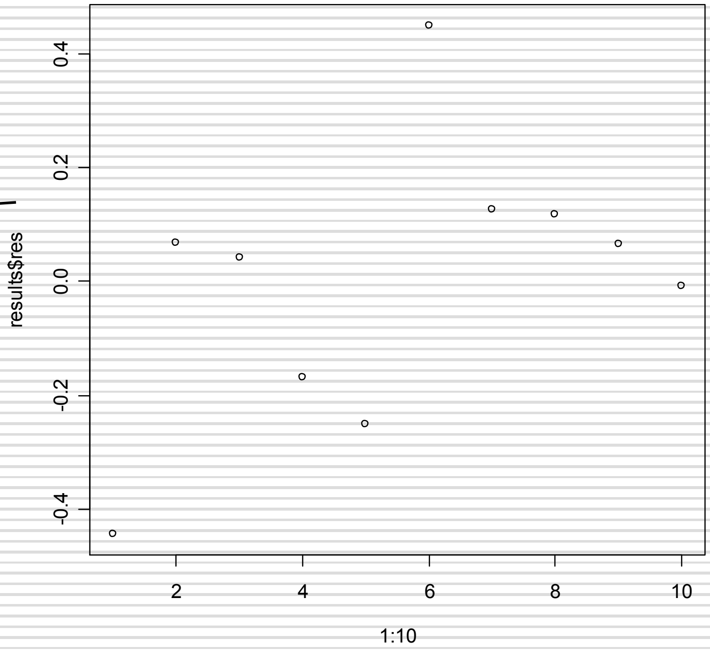
ομοσκεδαστικότητα



Παράδειγμα γενικού γραμμικού μοντέλου στην R

```
> plot(1:10,results$res)
```

ανεξαρτησία υπολοίπων



Παράδειγμα γενικού γραμμικού μοντέλου στην R

- Παρατηρούμε λοιπόν ότι

$$\hat{y} = 10.30 - 0.27x_1 + 8.49x_2.$$

- Ακόμα παρατηρούμε ότι ο συντελεστής προσδιορισμού $R^2 = 0.9$, έχουμε δηλαδή σχεδόν τέλεια προσαρμογή του μοντέλου, ενώ ο διορθωμένος είναι 0.87 επίσης πολύ υψηλός, ενώ

$$s_{y|x} = 0.27 \text{ (αρκετά μικρό).}$$

Παράδειγμα γενικού γραμμικού μοντέλου στην R

- Στον έλεγχο για το b_1 βλέπουμε ότι η P-τιμή δεν είναι πολύ μικρή οπότε σε ε.σ. 5% δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι αύξηση κατά μια μονάδα της X_1 δεν σημαίνει μεταβολή της αναμενόμενης τιμής της Y (υπό την προϋπόθεση ότι η X_2 παραμένει σταθερή).
- Στον έλεγχο για το b_2 βλέπουμε ότι η P-τιμή είναι πολύ μικρή οπότε πράγματι αύξηση κατά μια μονάδα της X_2 σημαίνει και μεταβολή της αναμενόμενης τιμής της Y (υπό την προϋπόθεση ότι η X_1 παραμένει σταθερή). Πιο συγκεκριμένα αύξηση του ειδικού βάρους των δοκών κατά 1 μονάδα σημαίνει αύξηση της αναμενόμενης αντοχής τους κατά 8.49, υπό την προϋπόθεση βέβαια ότι η περιεκτικότητα σε νερό παραμένει σταθερή.
- Τέλος η σταθερά φαίνεται από την p-τιμή του αντίστοιχου ελέγχου να είναι στατιστικά διάφορη του μηδενός. Πιο συγκεκριμένα η αναμενόμενη αντοχή των δοκών με μηδενική περιεκτικότητα σε νερό και μηδενικό ειδικό βάρος είναι 10.30. Φυσικά η συγκεκριμένη δήλωση δεν έχει ερμηνεία!
- Τέλος από το F-test καταλήγουμε να απορρίψουμε την υπόθεση ότι $b_1=b_2=0$, δηλαδή μπορούμε να εξηγήσουμε την μεταβλητότητα της αντοχής των δοκών από το γενικό γραμμικό μοντέλο.

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

- **Επιλογή Μεταβλητών:** Ένα πρόβλημα που αρκετές φορές αντιμετωπίζουμε είναι η επιλογή των κατάλληλων ανεξάρτητων μεταβλητών που θα χρησιμοποιήσουμε στο μοντέλο μας. Όταν στην διάθεσή μας έχουμε στοιχεία από πολλές επεξηγηματικές μεταβλητές δεν σημαίνει κατά ανάγκη ότι στο τελικό μοντέλο θα πρέπει να χρησιμοποιηθούν όλες αυτές. Υπάρχουν πολλά κριτήρια επιλογής μεταβλητών, η παρουσίαση των οποίων ξεφεύγει από τα όρια αυτών των σημειώσεων.

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

- Συγχυτικοί **Παράγοντες:** **Συγχυτικός παράγοντας** (*confounder*) ονομάζεται μια μεταβλητή, έστω Ζ, η οποία διαστρεβλώνει τη σχέση μεταξύ της μεταβλητής απόκρισης Υ και μιας επεξηγηματικής μεταβλητής Χ. Ο λόγος αυτής της διαστρέβλωσης είναι ότι ο συγχυτικός παράγοντας Ζ σχετίζεται και με την επεξηγηματική μεταβλητή Χ αλλά και με την μεταβλητή απόκρισης Υ. Άρα μπορεί η επεξηγηματική μεταβλητή Χ να είναι στατιστικά σημαντική (δηλαδή αύξησή της κατά μια μονάδα πράγματι να προκαλεί μεταβολή στην αναμενόμενη τιμή της μεταβλητής απόκρισης Υ) στο απλό γραμμικό μοντέλο $E(Y|X)=a+bX$, αλλά όταν στο μοντέλο προσθέσουμε και την μεταβλητή Ζ η επίδρασή της να εξαλείφεται και η Ζ να είναι αυτή που περιγράφει την μεταβλητότητα της Υ.

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

Ως παράδειγμα θεωρήστε τα ακόλουθα δεδομένα τα οποία μας δίνουν το ποσοστό των ενηλίκων που χρησιμοποιούν το διαδίκτυο (Internet), το Ακαθάριστο Εγχώριο Προϊόν (GDP) και το μέσο αριθμό παιδιών ανά ενήλικη γυναίκα (Fertility), για 39 χώρες.

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

INTERNET	GDP	FERTILITY
0.65	6.09	2.8
10.08	11.32	2.4
37.14	25.37	1.7
38.7	26.73	1.3
31.04	25.52	1.7
4.66	7.36	2.2
46.66	27.13	1.5
20.14	9.19	2.4
2.57	4.02	1.8
42.95	29.00	1.8
0.93	3.52	3.3
43.03	24.43	1.7
26.38	23.99	1.9

INTERNET	GDP	FERTILITY
37.36	25.35	1.4
13.21	17.44	1.3
0.68	2.84	3.0
1.56	6.00	2.3
23.31	32.41	1.9
27.66	19.79	2.7
38.42	25.13	1.3
27.31	8.75	2.9
3.62	8.43	2.5
49.05	27.19	1.7
46.12	19.16	2.0
0.10	0.85	5.4
46.38	29.62	1.8

INTERNET	GDP	FERTILITY
0.34	1.89	5.1
2.56	3.84	3.2
2.93	7.10	1.1
1.34	13.33	4.5
6.49	11.29	2.6
18.27	20.15	1.2
51.63	24.18	1.6
30.70	28.10	1.4
6.04	5.89	2.4
32.96	24.16	1.6
50.15	34.32	2.1
1.24	2.07	2.3
0.09	0.79	7.0

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

```
> summary(lm(data$INTERNET~data$FERTILITY))

Call:
lm(formula = data$INTERNET ~ data$FERTILITY)

Residuals:
    Min      1Q  Median      3Q     Max
-28.662 -13.890   1.424  10.436  26.727

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      40.577      5.451   7.444 7.32e-09 ***
data$FERTILITY   -8.169      2.035  -4.013  0.00028 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.62 on 37 degrees of freedom
Multiple R-squared: 0.3033,     Adjusted R-squared: 0.2845
F-statistic: 16.11 on 1 and 37 DF,  p-value: 0.0002803
```

```
>summary(lm(data$INTERNET~data$FERTILITY+data$GDP))

Call:
lm(formula = data$INTERNET ~ data$FERTILITY + data$GDP)

Residuals:
    Min      1Q  Median      3Q     Max
-23.155  -4.758   0.454   2.761  20.061

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.1970     5.6922  -0.562    0.578
data$FERTILITY  -0.1180     1.4399  -0.082    0.935
data$GDP         1.5392     0.1692   9.099 7.28e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.718 on 36 degrees of freedom
Multiple R-squared: 0.7889,     Adjusted R-squared: 0.7771
F-statistic: 67.25 on 2 and 36 DF,  p-value: 6.957e-13
```

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

- Παρατηρήστε πως η επίδραση της μεταβλητής Fertility στο γραμμικό μοντέλο εξαλείφεται όταν προσθέσουμε τον συγχυτικό παράγοντα GDP.
- Παρατηρήστε πως σχετίζονται μεταξύ τους οι 3 αυτές μεταβλητές.

```
> cor(data)
             INTERNET        GDP  FERTILITY
INTERNET    1.0000000  0.8881536 -0.5507279
GDP         0.8881536  1.0000000 -0.6145094
FERTILITY  -0.5507279 -0.6145094  1.0000000
```

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

- Πολυσυγγραμικότητα. Δεν θα πρέπει στο μοντέλο παλινδρόμησης να συμπεριλαμβάνουμε επεξηγηματικές μεταβλητές με **υψηλή συσχέτιση** μεταξύ τους, διότι τότε τα αποτελέσματα δεν είναι έγκυρα. Ως παράδειγμα θεωρήστε τα ακόλουθα δεδομένα τα οποία μας δίνουν το ύψος 25 ανδρών σε inches (Height) καθώς και το μήκος σε cm του αριστερού (LeftFoot) καθώς και δεξιού τους ποδιού (RtFoot). Σκοπός είναι να προσαρμόσουμε ένα γραμμικό μοντέλο με μεταβλητή απόκρισης Height και επεξηγηματικές μεταβλητές LeftFoot και RtFoot.

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

Height	LeftFoot	RtFoot
69.0	27.0	26.5
79.0	29.0	27.5
75.0	31.0	32.0
69.0	25.5	25.5
65.0	23.5	23.0
79.0	28.0	28.0
72.0	28.5	28.5
69.5	27.0	27.0
73.0	30.6	31.4
71.5	27.4	28.5
69.5	27.0	27.0
73.0	28.5	27.5
71.0	29.0	27.0

Height	LeftFoot	RtFoot
74.0	29.0	30.0
75.0	28.0	29.0
71.0	27.0	27.5
72.0	26.5	27.5
66.0	25.5	26.0
71.0	29.0	28.0
67.0	27.2	27.0
71.0	29.0	28.5
72.0	28.0	28.0
72.0	28.5	29.0
73.5	29.0	30.0
73.0	23.5	24.0

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

```
>summary(lm(data$Height~data$LeftFoot+data$RtFoot))

Call:
lm(formula = data$Height ~ data$LeftFoot + data$RtFoot)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1330 -1.6460 -0.5014  0.6246  6.9920

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    43.9334     8.9983   4.882    7e-05 ***
data$LeftFoot   0.6379     0.7730   0.825    0.418
data$RtFoot     0.3647     0.7096   0.514    0.612
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.905 on 22 degrees of freedom
Multiple R-squared: 0.3072,  Adjusted R-squared: 0.2442
F-statistic: 4.877 on 2 and 22 DF,  p-value: 0.01765
```

```
> summary(lm(data$Height~data$LeftFoot))

Call:
lm(formula = data$Height ~ data$LeftFoot)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2327 -2.0302 -0.5309  0.4698  6.9684

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.0701     8.8493   4.980  4.9e-05 ***
data$LeftFoot   0.9986     0.3189   3.131  0.00469 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.859 on 23 degrees of freedom
Multiple R-squared: 0.2989,     Adjusted R-squared: 0.2684
F-statistic: 9.804 on 1 and 23 DF,  p-value: 0.004689
```

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

```
> summary(lm(data$Height~data$RtFoot))

Call:
lm(formula = data$Height ~ data$RtFoot)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1460 -1.5424 -0.5241  0.2686  7.5095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.8408     8.2224   5.697 8.43e-06 ***
data$RtFoot   0.8964     0.2955   3.033  0.00591 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.885 on 23 degrees of
        freedom
Multiple R-squared: 0.2857,     Adjusted R-squared:
        0.2547
F-statistic: 9.201 on 1 and 23 DF,  p-value: 0.00591
```

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

- Παρατηρήστε πως όταν συμπεριλαμβάνουμε και τις δύο επεξηγηματικές μεταβλητές στο μοντέλο (γενικό γραμμικό μοντέλο) καμία δεν είναι στατιστικά σημαντική, ενώ αντίθετα κάθε μία εξ αυτών έχει πολύ μεγάλη επίδραση στην μεταβλητή απόκρισης (όπως προφανώς αναμενόταν) όπως παρατηρούμε από τα απλά γραμμικά μοντέλα.
- Το παραπάνω φαινόμενο οφείλεται στην υψηλή συσχέτιση των δύο επεξηγηματικών μεταβλητών.

```
> cor(data$RtFoot,data$LeftFoot)
[1] 0.9078141
```

Διάφορα θέματα στο γενικό γραμμικό μοντέλο

- Πρόβλεψη: Οι προβλέψεις είναι σωστές και αξιόπιστες μόνο για τις τιμές του X που είναι σχετικά κοντά με αυτές που έχουμε παρατηρήσει. Στο παράδειγμα με τις ξύλινες δοκούς δεν θα ήταν λογικό να εκτιμούσαμε την αντοχή των δοκών με περιεκτικότητα σε νερό 30.

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Μέχρι τώρα οι επεξηγηματικές μεταβλητές του μοντέλου παλινδρόμησης θεωρήσαμε ότι είναι ποσοτικές τυχαίες μεταβλητές. Ωστόσο, μερικές φορές έχουμε επεξηγηματικές μεταβλητές που αποδίδουν έναν παράγοντα με δύο ή περισσότερες κατηγορίες, όπως π.χ. το φύλο ή την φυσική κατάσταση ενός ατόμου (κακή, μέτρια, καλή και άριστη). Ας υποθέσουμε παραδείγματος χάρη, ότι θέλουμε να δημιουργήσουμε ένα μοντέλο γραμμικής παλινδρόμησης με μεταβλητή απόκρισης Y το καθαρό εισόδημα και επεξηγηματική μεταβλητή X το επίπεδο εκπαίδευσης (δημοτικό, γυμνάσιο, λύκειο, πανεπιστήμιο). Τότε δημιουργούμε τις λεγόμενες εικονικές μεταβλητές ή ψευδομεταβλητές, οι οποίες είναι (m-1) στον αριθμό, όπου m είναι ο αριθμός των κατηγοριών της κατηγορικής μεταβλητής X, και λαμβάνουν τις τιμές 0 ή 1. Στο παράδειγμα με το εισόδημα και το επίπεδο εκπαίδευσης (4 κατηγορίες), ορίζουμε τις εξής 3 εικονικές μεταβλητές:

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

$$X_1 = \begin{cases} 1, & \text{δημοτικό} \\ 0, & \text{διαφορετικά} \end{cases} \quad X_2 = \begin{cases} 1, & \text{γυμνάσιο} \\ 0, & \text{διαφορετικά} \end{cases} \quad X_3 = \begin{cases} 1, & \text{λύκειο} \\ 0, & \text{διαφορετικά} \end{cases}$$

και εκτιμούμε με την μέθοδο ελαχίστων τετραγώνων τους συντελεστές του γραμμικού μοντέλου

$$E\left[Y \middle| X_1, X_2, X_3\right] = a + b_1 x_1 + b_2 x_2 + b_3 x_3.$$

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Ισοδύναμα με το παραπάνω γραμμικό μοντέλο είναι σαν να έχουμε τα ακόλουθα 4 γραμμικά μοντέλα:

$$E(Y \mid X_1, X_2, X_3) = \begin{cases} a, & \text{πανεπιστήμιο} \\ a + b_1, & \text{δημοτικό} \\ a + b_2, & \text{γυμνάσιο} \\ a + b_3, & \text{λύκειο} \end{cases}.$$

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Η κατηγορία για την οποία δεν ορίσαμε ψευδομεταβλητή ονομάζεται κατηγορία αναφοράς (*reference category*), στο παράδειγμα μας η κατηγορία αυτή είναι η πανεπιστημιακή μόρφωση, και η επιλογή της εξαρτάται συχνά από το υπό μελέτη πρόβλημα. Αν στο παράδειγμά μας θέλουμε να εξετάσουμε αν η πανεπιστημιακή μόρφωση αυξάνει κατά μέσο όρο το εισόδημα, τότε είναι φυσικό να θέσουμε ως κατηγορία αναφοράς την εν λόγω κατηγορία.
- Οι παράμετροι του παραπάνω γραμμικού μοντέλου μπορούν να ερμηνευτούν αν θεωρήσουμε τις πιθανές τιμές των εικονικών μεταβλητών. Έτσι η παράμετρος a στο παραπάνω παράδειγμα εκφράζει το μέσο καθαρό εισόδημα όταν $X_1 = X_2 = X_3 = 0$, άρα όταν η μόρφωση του ατόμου είναι πανεπιστημιακή. Η παράμετρος b_1 αντιπροσωπεύει την διαφορά μεταξύ του μέσου καθαρού εισοδήματος ατόμων με μόρφωση δημοτικού από το μέσο καθαρό εισόδημα ατόμων με πανεπιστημιακή μόρφωση. Ανάλογα ερμηνεύονται και οι υπόλοιποι συντελεστές.
- Τα γραμμικά μοντέλα με μία κατηγορική επεξηγηματική μεταβλητή είναι ισοδύναμα με τα μοντέλα Ανάλυσης Διασποράς με έναν Παράγοντα (*One Way ANOVA*). Αν η κατηγορική μεταβλητή είναι δίτιμη τα συμπεράσματα του γραμμικού μοντέλο είναι ισοδύναμα με αυτά του one sample t-test που είδαμε στο προηγούμενο κεφάλαιο.

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Παράδειγμα: Μετρήσεις του ετήσιου μισθού Y (σε χιλιάδες €) των υπαλλήλων μιας εταιρείας, συναρτήσει της εμπειρίας τους X_1 (σε έτη υπηρεσίας) και του μορφωτικού τους επιπέδου X_2 (1 = απολυτήριο λυκείου, 2 = απόφοιτος ΑΕΙ και 3 = μεταπτυχιακός τίτλος) δίνονται από τον πίνακα της επόμενης διαφάνειας. Να εκτιμηθούν και να ερμηνευτούν με την μέθοδο ελαχίστων τετραγώνων οι συντελεστές του μοντέλου

$$E(Y \mid X_1, X_{22}, X_{23}) = a + b_1 x_1 + b_2 x_{22} + b_3 x_{23}$$

$$\text{όπου} \quad X_{22} = \begin{cases} 1, & \text{απόφοιτος ΑΕΙ} \\ 0, & \text{διαφορετικά} \end{cases} \qquad X_{23} = \begin{cases} 1, & \text{μεταπτυχιακός τίτλος} \\ 0, & \text{διαφορετικά} \end{cases}$$

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

Y	25	22	20	32	35	28	19	17	20	26	28	34	40	42	28	19
X_1	2	1	2	5	7	3	2	1	2	2	5	5	7	6	3	1
X_2	2	2	1	3	2	3	1	1	2	2	2	3	3	3	2	1

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Έστω τα δεδομένα είναι αποθηκευμένα στο αρχείο data.txt (στον φάκελο από όπου τρέχουμε την R) υπό μορφή πίνακα διάστασης 16×3.

```
> data<-matrix(scan("data.txt"), ncol=16, byrow=T)
Read 48 items
> data
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]   25   22   20   32   35   28   19   17   20    26    28    34    40    42
[2,]    2    1    2    5    7    3    2    1    2     2     5     5     7     6
[3,]    2    2    1    3    2    3    1    1    2     2     2     3     3     3
     [,15] [,16]
[1,]    28    19
[2,]     3     1
[3,]     2     1
> y<-data[1,]
> x1<-data[2,]
> x2<-data[3,]
```

Διαβάζουμε τα δεδομένα από το αρχείο

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

```
> x2<-as.factor(x2)
> x2
[1] 2 2 1 3 2 3 1 1 2 2 2 3 3 3 2 1
Levels: 1 2 3
> results<-lm(y~x1+x2)
> summary(results)
Call:
lm(formula = y ~ x1 + x2)
Residuals:
    Min      1Q  Median      3Q     Max 
-3.5729 -1.1971 -0.1892  1.4296  4.9010 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15.1894     1.3709  11.080 1.17e-07 ***
x1            2.3737     0.4058   5.850 7.84e-05 ***
x22           3.6360     1.6780   2.167   0.0511 .  
x23           7.6672     2.2294   3.439   0.0049 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 2.457 on 12 degrees of freedom
Multiple R-squared: 0.9187,     Adjusted R-squared: 0.8983 
F-statistic: 45.18 on 3 and 12 DF,  p-value: 8.193e-07
```

μετατρέπουμε την μεταβλητή σε κατηγορική

Η R παίρνει αυτόματα την 1η κατηγορία ως κατηγορία αναφοράς

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

```
> model.matrix(results)
   (Intercept) x1 x22 x23
1            1  2   1   0
2            1  1   1   0
3            1  2   0   0
4            1  5   0   1
5            1  7   1   0
6            1  3   0   1
7            1  2   0   0
8            1  1   0   0
9            1  2   1   0
10           1  2   1   0
11           1  5   1   0
12           1  5   0   1
13           1  7   0   1
14           1  6   0   1
15           1  3   1   0
16           1  1   0   0
```

βλέπουμε τις εικονικές μεταβλητές που δημιουργήθηκαν

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Παρατηρούμε λοιπόν ότι: (α) ο αναμενόμενος μισθός ενός υπαλλήλου της εν λόγω εταιρείας με απολυτήριο λυκείου χωρίς προϋπηρεσία είναι €15190, (β) η αύξηση ενός έτους στα χρόνια εμπειρίας μεταφράζεται σε αύξηση του αναμενόμενου μισθού κατά €2370 για υπαλλήλους ανεξαρτήτως πανεπιστημιακής μόρφωσης, (γ) οι απόφοιτοι ΑΕΙ έχουν αναμενόμενο μισθό κατά €3640 υψηλότερο σε σχέση με τους συναδέλφους τους με τα ίδια χρόνια εμπειρίας που έχουν απολυτήριο λυκείου και (δ) οι κάτοχοι μεταπτυχιακών τίτλων έχουν αναμενόμενο μισθό κατά €7670 υψηλότερο σε σχέση με τους συναδέλφους τους με τα ίδια χρόνια εμπειρίας που έχουν απολυτήριο λυκείου.

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Παρατηρούμε λοιπόν ότι

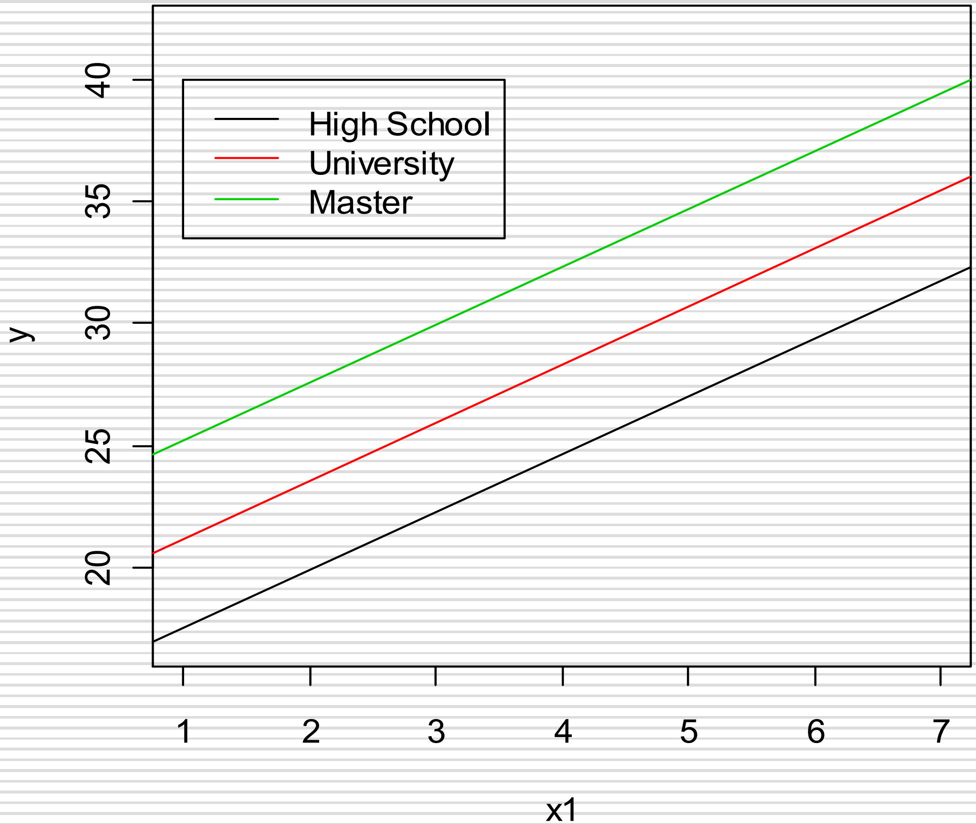
$$E(Y \mid X_1, X_{22}, X_{23}) = 15.19 + 2.37x_1 + 3.64x_{22} + 7.67x_{23}$$

- Ισοδύναμα

$$E(Y \mid X_1, X_{22}, X_{23}) = \begin{cases} 15.19 + 2.37X_1, \ \text{απολυτήριο λυκείου} \\ (15.19 + 3.64) + 2.37X_1, \ \text{απόφοιτος ΑΕΙ} \\ (15.19 + 7.67) + 2.37X_1, \ \text{μεταπτυχιακός τίτλος} \end{cases}$$

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

```
> plot(x1,y,type='n')
> abline(15.19,2.37, col=1)
> abline(18.83,2.37, col=2)
> abline(22.86,2.37, col=3)
> legend(1,40, lty=1,
    col=1:3,legend=c("High School",
    "University", "Master"))
```



Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

- Αν θέλαμε να χρησιμοποιήσετε άλλη κατηγορία αναφοράς, αλλάζουμε την σειρά των κατηγοριών της μεταβλητής, έτσι ώστε η πρώτη να είναι αυτή που θέλουμε να είναι κατηγορία αναφοράς. Π.χ. αν θέλουμε η δεύτερη κατηγορία (απόφοιτος ΑΕΙ) να είναι η κατηγορία αναφοράς κάνουμε τα εξής:

```
> x2 <- factor(x2, levels = c(2, 1, 3))
> x2
[1] 2 2 1 3 2 3 1 1 2 2 2 3 3 3 2 1
Levels: 2 1 3
> lm(y~x1+x2)
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)          x1         x21         x23
     18.825       2.374      -3.636       4.031
```

Εικονικές Μεταβλητές στο γενικό γραμμικό μοντέλο

```
> model.matrix(lm(y~x1+x2))
   (Intercept) x1 x21 x23
1            1  2   0   0
2            1  1   0   0
3            1  2   1   0
4            1  5   0   1
5            1  7   0   0
6            1  3   0   1
7            1  2   1   0
8            1  1   1   0
9            1  2   0   0
10           1  2   0   0
11           1  5   0   0
12           1  5   0   1
13           1  7   0   1
14           1  6   0   1
15           1  3   0   0
16           1  1   1   0
```

Επίλογος

- Αν η μεταβλητή απόκρισης Y είναι κατηγορική μεταβλητή, τότε το γραμμικό μοντέλο δεν είναι πλέον το κατάλληλο, και χρησιμοποιούμε γενικευμένα γραμμικά μοντέλα, η παρουσίαση των οποίων ξεφεύγει από τους σκοπούς αυτών των σημειώσεων.