Εφαρμοσμένη Στατιστική

Δημήτριος Μπάγκαβος

Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών
Πανεπιστήμιο Κρήτης

20 Μαρτίου 2018

Ανάλυση Παλινδρόμησης: Γενικά.

- Με την ανάλυση παλινδρόμησης εξετάζουμε τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης.
- Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε δύο είδη μεταβλητών: τις ανεξάρτητες ή επεξηγηματικές και τις εξαρτημένες.
- Στην πράξη ανεξάρτητη μεταβλητή X είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή, να καθορίσουμε τις τιμές της (π.χ. το ύψος της διαφημιστικής δαπάνης ενός προϊόντος, ο αριθμός των λειτουργούντων ταμείων σε ένα υποκατάστημα τραπέζης, η ποσότητα λιπάσματος, η θερμοκρασία επεξεργασίας ενός προϊόντος).
- Εξαρτημένη μεταβλητή Y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές (π.χ. η ζήτηση ενός προϊόντος, ο χρόνος αναμονής των πελατών ενός υποκαταστήματος τραπέζης, η απόδοση μιας καλλιέργειας, η αντοχή ενός υλικού).
- Όταν η μεταβολή στην τιμή μίας τυχαίας μεταβλητής X συνεπάγεται και μεταβολή στην τιμή της Y, τότε μιλάμε για συσχέτιση μεταξύ των δύο αυτών μεταβλητών.

Ανάλυση Παλινδρόμησης: Γενικά.

- Η συσχέτιση μεταξύ των X, Y μπορεί να είναι δυο ειδών ανάλογα με το είδος του προβλήματος που μελετάμε.
  - Ντετερμινιστική σχέση: οι X, Y συνδέονται με μια σχέση της μορφής $Y = f(X)$ που δεν υπόκειται σε σφάλματα.
  - Δηλαδή για κάθε τιμή της X μπορούμε να προβλέψουμε ακριβώς την τιμή της Y.
  - Για παράδειγμα, το ποσό που καταθέτει κάποιος στο Ταμιευτήριο και ο τόκος που παίρνει για το ποσό αυτό, συνδέονται με ντετερμινιστική σχέση.
  - Σε αυτές τις περιπτώσεις τα σημεία του διαγράμματος διασποράς βρίσκονται όλα πάνω στην καμπύλη που έχει εξίσωση $Y = f(X)$
  - Όσες φορές και αν επαναλάβουμε το πείραμα θέτοντας το X στο ίδιο επίπεδο $X = x_i$ θα παίρνουμε πάντα την ίδια τιμή για το Y.
  - Στοχαστική – στατιστική σχέση: Στην περίπτωση αυτή, αν επαναλάβουμε το πείραμα πολλές φορές θέτοντας το X στο ίδιο επίπεδο $X = x_i$ τότε στην τιμή x_i της X δεν αντιστοιχεί μια μόνο τιμή y_i της Y αλλά, γενικά, αντιστοιχεί ένα πλήθος διαφορετικών τιμών της Y.
  - Για παράδειγμα, αν X είναι η τιμή ενός προϊόντος και Y είναι η ζήτησή του, η σχέση των Y, X είναι στοχαστική γιατί η τιμή καθορίζει τη ζήτηση, όχι όμως επακριβώς.

Ανάλυση Παλινδρόμησης: Συσχέτιση.

- Εδώ μας ενδιαφέρει αποκλειστικά η μελέτη στοχαστικών φαινομένων. Μας ενδιαφέρει να βρούμε αν:
  - Υπάρχει σχέση ανάμεσα σε δύο (ή περισσότερες) μεταβλητές,
  - Αν υπάρχει σχέση ποια η φύση της σχέσης αυτής:
    - Θετικά συσχετισμένες;
    - Αρνητικά συσχετισμένες;
  - Αν δεν υπάρχει σχέση μεταξύ των X και Y τότε λέμε ότι οι δυο μεταβλητές είναι ασυσχέτιστες.

  - Η παλινδρόμηση ποσοτικοποιεί την εξάρτηση, όταν αυτή υπάρχει.
  - Είναι σημαντικό η παλινδρόμηση να εφαρμόζεται αφού έχουμε διαπιστώσει ότι υπάρχει συσχέτιση μεταξύ των μεταβλητών, αλλιώς τα αποτελέσματα της ανάλυσης θα είναι λάθος.

- Η συσχέτιση δυο τ.μ. μπορεί να οριστεί ως προς διάφορα μεγέθη:
  1. Ως προς τις τιμές των μεταβλητών (συσχέτιση Pearson: κυρίως για συνεχή, κανονικά κατανεμημένα δεδομένα),
  2. Ως προς τις διατάξεις των τιμών (συσχέτιση Spearman: κυρίως για μεταβλητές διάταξης - επίσης δεν είναι απαραίτητο να έχουμε κανονικά ή συνεχή δεδομένα)
  3. Ως προς σύνολα δεδομένων (συντελεστής Kendall)

Ανάλυση Παλινδρόμησης: Συσχέτιση.

- Έστω $X_i, Y_i, i = 1, \ldots, n$ οι τιμές των X, Y αντίστοιχα.
- Ο συντελεστής συσχέτισης του Pearson ορίζεται από την

$$\rho = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left(\sum_{i=1}^{n} X_i - \bar{X}\right)\left(\sum_{i=1}^{n} Y_i - \bar{Y}\right)} \tag{1}$$

- Έστω $X_{(i)}, Y_{(i)}, i = 1, \ldots, n$ οι διατεταγμένες τιμές (από την μικρότερη στη μεγαλύτερη) X_i, Y_i.
- Ο συντελεστής συσχέτισης του Spearman δίνεται από την

$$\rho_s = \frac{\sum_{i=1}^{n} \left(X_{(i)} - \bar{X}\right)\left(Y_{(i)} - \bar{Y}\right)}{\left(\sum_{i=1}^{n} X_{(i)} - \bar{X}\right)\left(\sum_{i=1}^{n} Y_{(i)} - \bar{Y}\right)} \tag{2}$$

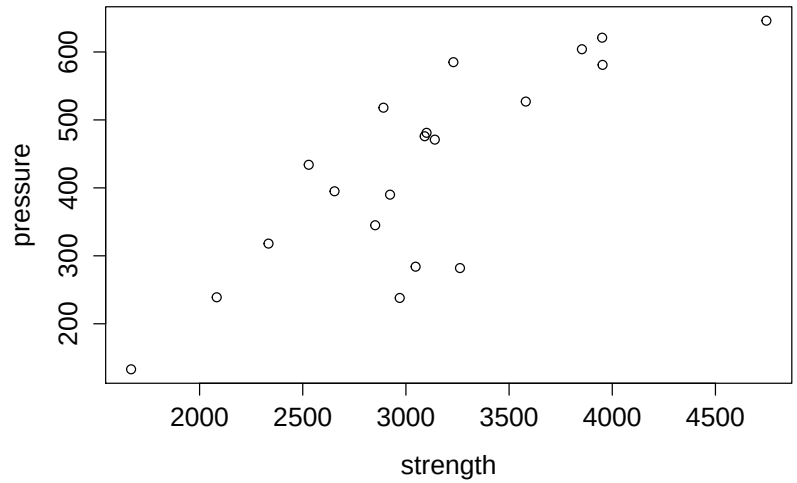
- Ο συντελεστής του Spearman χρησιμοποιείται όταν οι παραδοχές του συντελεστή του Pearson δεν πληρούνται, δηλαδή όταν τα δεδομένα δεν είναι συνεχή ή κανονικά κατανεμημένα.
- Επίσης προτιμάται όταν οι τιμές μιας μεταβλητής είναι έντονα ασύμμετρες.

Ανάλυση Παλινδρόμησης: Συσχέτιση.

- Και οι δύο συντελεστές μετράνε γραμμική εξάρτηση, δηλαδή το κατά πόσο καλά μπορεί να προσαρμοστεί μια ευθεία στα δεδομένα.
- Πάντα έχουμε $-1 \leq \rho \leq 1$. Συγκεκριμένα:
  - $\rho > 0$: θετική συσχέτιση: αύξηση της τιμής της X συνεπάγεται αύξηση της τιμής της Y
  - $\rho < 0$: αρνητική (αντιστρόφως ανάλογη) συσχέτιση: αύξηση της τιμής της X συνεπάγεται μείωση της τιμής της Y
  - $\rho = \pm 1$: Η σχέση μεταξύ X και Y δεν είναι τυχαία (οι τιμές της X καθορίζουν πλήρως τις τιμές της Y)
- Αν η συσχέτιση είναι καμπυλόγραμμη, οι συντελεστές Pearson ή Spearman μπορεί να είναι παραπλανητικοί.
- Γι΄ αυτό και το πρώτο συμπέρασμα για την συσχέτιση η όχι τυχαίων μεταβλητών προέρχεται από ένα γράφημα (scatterplot) μεταξύ των τιμών τους (γράφημα διασποράς).
- Ένα γράφημα των $(X_i, Y_i), i = 1, \ldots, n$ περιμένουμε να αποκαλύψει αν φαίνεται κάποια σχέση μεταξύ των δυο μεταβλητών.
- Αν μπορούμε να διακρίνουμε κάποια πατέντα στο γράφημα, αυτή προέρχεται από τη σχέση των X, Y.
- Σαν παράδειγμα χρησιμοποιούμε τις δύο μεταβλητές από το αρχείο Graph1.csv που περιγράφουν μετρήσεις πίεσης και αντοχής σε κάποιο υλικό.

Ανάλυση Παλινδρόμησης: Συσχέτιση.

- Από το γράφημα φαίνεται ξεκάθαρα ότι υπάρχει μία γραμμική σχέση μεταξύ πίεσης και αντοχής:
- όσο πιο πολύ αυξάνει η πίεση τόσο πιο μεγάλες είναι οι μετρήσεις αντοχής.



- Για να υπολογίσουμε το συντελεστή συσχέτισης μεταξύ των 2 μεταβλητών χρησιμοποιούμε την εντολή `cor(Graph1, method="pearson")` αν θέλουμε να υλοποιήσουμε την (1), ενώ με `method="spearman"` υλοποιούμε την (2).
- Στην περίπτωση μας βλέπουμε ότι η συσχέτιση μεταξύ της `pressure` και της `strength` είναι ίση με 0.807 το οποίο επιβεβαιώνει την οπτική εντύπωση του γραφήματος.
- Σημείωση: Ποτέ δεν πρέπει να παρουσιάζουμε ένα συντελεστή συσχέτισης χωρίς να εξετάζουμε το γράφημα διασποράς.
- Ο λόγος είναι ότι από το γράφημα ελέγχουμε για τυχόν προβλήματα όπως μη γραμμικές σχέσεις και έντονα αποκλίνουσες τιμές.

Ανάλυση Παλινδρόμησης: Συσχέτιση.

- Ο συντελεστής του Kendall (Kendall's tau) υπολογίζεται με την `method="kendall"` στην `cor()` και ορίζεται από την

$$\tau = 2\frac{N_c - N_d}{n(n-1)} \tag{3}$$

όπου n είναι ο αριθμός των παρατηρήσεων και N_c και N_d είναι οι αριθμοί των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων, αντίστοιχα.

- Δύο παρατηρήσεις, έστω (X_j, Y_j) και (X_k, Y_k), ονομάζονται εναρμονισμένες, αν και τα δύο μέλη της μίας παρατήρησης είναι μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης, δήλ. αν $X_j > X_k$ (αντίστοιχα, $X_j < X_k$), τότε $Y_j > Y_k$ (αντίστοιχα, $Y_j < Y_k$).
- Αντίστοιχα, δύο παρατηρήσεις λέγονται μη εναρμονισμένες παρατηρήσεις, αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την διάταξη των δεύτερων μελών τους, δηλαδή, αν $X_j > X_k$ (αντίστοιχα, $X_j < X_k$), τότε $Y_j < Y_k$ (αντίστοιχα, $Y_j > Y_k$).
- Στα χαρακτηριστικά του συντελεστή του Kendall περιλαμβάνεται ότι

Ανάλυση Παλινδρόμησης: Συσχέτιση.

1. Δεν κάνει κάποια υπόθεση σχετικά με τα δεδομένα,
2. Ενδείκνυται για μικρά σύνολα δεδομένων με πολλές ισοπαλίες στις βαθμίδες
3. Θεωρείται καλύτερος εκτιμητής της συσχέτισης που υπάρχει στον πληθυσμό
4. Γενικά δίνει μικρότερες συσχετίσεις από του Spearman

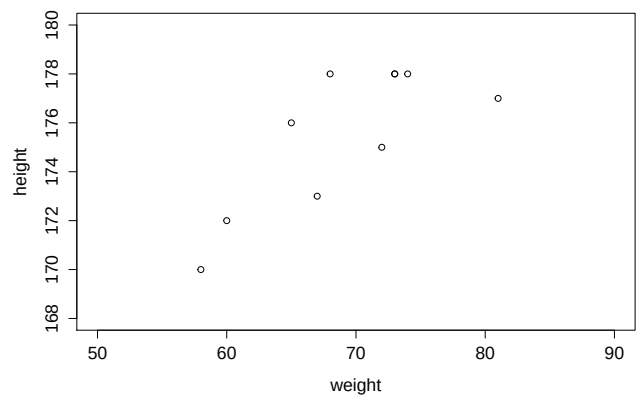
- Στο προηγούμενο παράδειγμα με τις μεταβλητές `pressure` και `strength` η `cor(Graph1, method="kendall")` δίνει συσχέτιση ίση με 0.589.
- Οι ακόλουθες μεταβλητές περιγράφουν το ύψος και το βάρος 10 παιδιών

```
weight<-c(58,60,67,72,65,81,73,74,73,68)
height<-c(170,172,173,175,176,177,178,178,178,178)
```

- Υπάρχει σχέση μεταξύ ύψους και βάρους;

```
cor(weight, height, method="pearson")
[1] 0.7775565
```

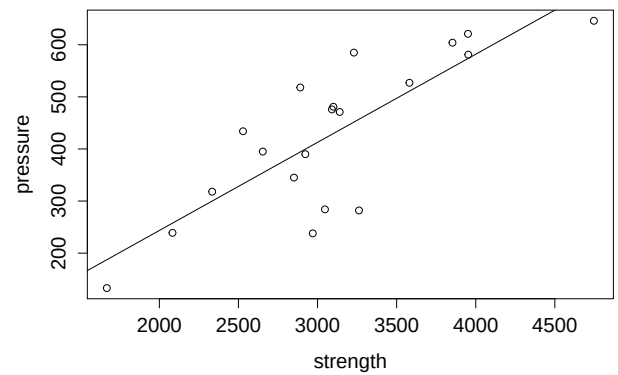
- Ερμηνεία: Υπάρχει μια σημαντική θετική σχέση μεταξύ βάρους και ύψους.
- Παιδιά με μικρότερο βάρος έχουν μικρότερο ύψος.



Ανάλυση παλινδρόμησης: μοντελοποίηση της συσχέτισης.

- Η ποσοτικοποίηση / εύρεση εξίσωσης που περιγράφει τη σχέση δύο μεταβλητών λέγεται ανάλυση παλινδρόμησης.
  - Στην πιο απλή περίπτωση, θέλουμε να προσδιορίσουμε την γραμμική σχέση μεταξύ μόλις δυο μεταβλητών.
  - Σε αυτή την περίπτωση έχουμε απλή γραμμική παλινδρόμηση ή το απλό γραμμικό μοντέλο.

- Τέτοια περίπτωση έχουμε όταν το γράφημα μεταξύ των X, Y μας δείχνει ότι θα περιγράφαμε ικανοποιητικά τη σχέση τους αν ορίζαμε μια ευθεία με βάση τα σημεία $(x_i, y_i), i = 1, \ldots, n$ όπως στη δίπλα εικόνα.

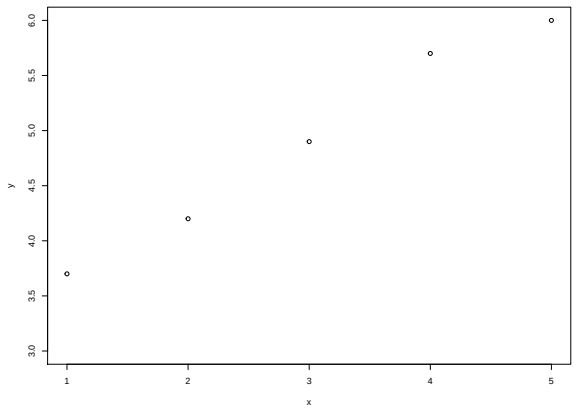


- Έχει πολύ μεγάλη σημασία να μπορούμε να προσδιορίσουμε σωστά ποια μεταβλητή εξαρτάται από την άλλη.
- Αυτό γιατί $Y = a + bX \neq X = a_1 + b_1 Y$. Δηλ. λάθος προσδιορισμός σημαίνει λάθος συμπεράσματα.
- Ο προσδιορισμός της ευθείας πρέπει να γίνει με τέτοιο τρόπο ώστε το άθροισμα των τετραγώνων των αποστάσεων της ευθείας από τα σημεία να είναι ελάχιστο.

Ανάλυση παλινδρόμησης: ευθεία ελαχίστων τετραγώνων.

- Έχουμε τα δεδομένα:

```
x<-c(1,2,3,4,5)
y<-c(3.7, 4.2, 4.9, 5.7, 6)
plot(x,y, ylim=c(3,6))
```



- Η σχέση μεταξύ των δύο μεταβλητών είναι σχεδόν γραμμική οπότε εφαρμόζουμε (προσεγγιστικά) το απλό γραμμικό μοντέλο $Y = \alpha + \beta X$.
- Η ακρίβεια της προσέγγισης εξαρτάται από το κατά πόσο το δ/μα $\varepsilon = Y - \alpha - \beta X$ μπορεί να ελαχιστοποιηθεί.
- Γενικά θεωρούμε n ζεύγη παρατηρήσεων $(X_i, Y_i), i = 1, \ldots, n$ και ψάχνουμε προσέγγιση της μορφής

$$Y_i = a + bX_i + \varepsilon_i \tag{4}$$

ώστε οι αποκλίσεις ε_i από τις παρατηρήσεις να είναι ελάχιστες.

- Αυτό επιτυγχάνεται με το να εκτιμήσουμε τα a, b ώστε το άθροισμα των τετραγώνων των αποκλίσεων

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \tag{5}$$

Ανάλυση παλινδρόμησης: ευθεία ελαχίστων τετραγώνων.

- να είναι ελάχιστο.
- Η ελαχιστοποίηση του αθροίσματος των ε_i δεν αποτελεί ασφαλές κριτήριο επιλογής διότι κάποια αρνητικά ε_i θα αναιρούν αντίστοιχες θετικές ποσότητες του αθροίσματος).
- Παραγωγίζοντας την (5) ως προς a και b και εξισώνοντας με μηδέν παίρνουμε τις ακόλουθες δύο εξισώσεις που ονομάζονται κανονικές εξισώσεις:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i,$$
$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \qquad (6)$$

- Λύνοντας το σύστημα των κανονικών εξισώσεων, παίρνουμε:

$$\hat{b} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2},$$
$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Ανάλυση παλινδρόμησης: Παράδειγμα 1.

Πρακτική άσκηση 1:

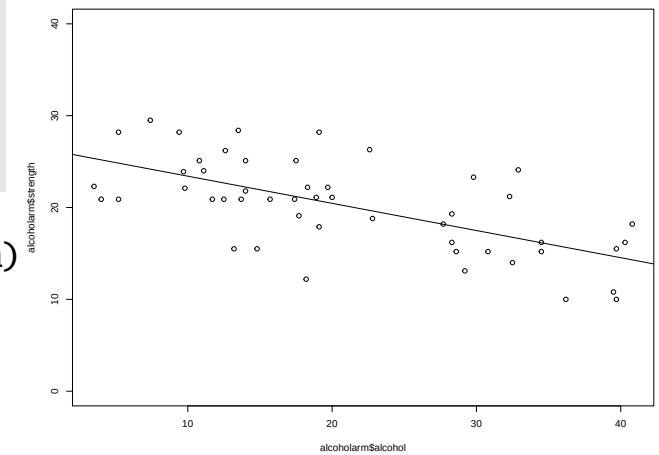
Θέλουμε να εξετάσουμε αν η κατανάλωση αλκοόλ συνδέεται γραμμικά με τη μυική δύναμη. Στα δεδομένα του αρχείου alcoholarm.txt στο moodle έχουμε ένα δείγμα 50 ανδρών για τους οποίους διαθέτουμε την κατανάλωση αλκοόλ και τη μυική δύναμη έκαστου.

- Οι εντολές με τις οποίες μπορούμε να εκτιμήσουμε τις παραμέτρους του γραμμικού μοντέλου, να απεικονίσουμε τα δεδομένα και την εκτιμώμενη ευθεία παλινδρόμησης είναι:

```
al.lm<-lm( strength ~ alcohol , data=
     alcoholarm)
plot(alcoholarm$alcohol, alcoholarm$
     strength, ylim=c(0,40))
abline(al.lm$coefficients[1], al.lm$
     coefficients[2])
```

```
Call:
lm(formula = strength ~ alcohol, data = alcoholarm)

Coefficients:
(Intercept)      alcohol
    26.3695      -0.2959
```



- Η προσαρμοσμένη ευθεία παλινδρόμησης δείχνει μια ελαφρά αρνητική σχέση μεταξύ κατανάλωσης αλκοόλ και μυικής δύναμης.

Ανάλυση παλινδρόμησης: ευθεία ελαχίστων τετραγώνων.

- Αν αντικαταστήσουμε τις τιμές των $\hat{a}, \hat{b}$ στην (4) παίρνουμε τις εκτιμώμενες τιμές $\hat{Y}_i$,

$$\hat{Y}_i = \bar{Y} - \hat{b}\bar{X} + \frac{n\sum_{i=1}^n X_i Y_i - (\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} X_i$$

$$= \bar{Y} + \frac{S_{XY}}{S_X^2}(X_i - \bar{X}) \qquad (7)$$

όπου

$$S_{XY} = n\sum_{i=1}^n X_i Y_i - \left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right),\ S_X^2 = n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2$$

- Από την (7) είναι ξεκάθαρο ότι η ευθεία της παλινδρόμησης διέρχεται από το σημείο $(\bar{X}, \bar{Y})$.
- Συγκρίνοντας την (4) με την $\hat{Y}_i = \hat{a} + \hat{b}X_i$ (ισοδύναμα με την (7)) βλέπουμε ότι η διαφορά τους σε κάθε παρατήρηση είναι ο όρος ε_i, οπότε γενικά $\hat{Y}_i \neq Y_i$.
- Η διαφορά $\varepsilon_i = Y_i - \hat{Y}_i$ είναι τυχαία μεταβλητή, $\varepsilon_i \sim N(0, \sigma^2)$.
- Οπότε, για κάθε παρατήρηση έχουμε

$$Y_i \sim N(a + bX_i, \sigma^2).$$

Ανάλυση παλινδρόμησης: υποθέσεις.

- Προηγουμένως υποθέσαμε ότι
  1. $\mathbb{E}\varepsilon_i = 0$
  2. $V(\varepsilon_i) = \sigma^2$ (ομοσκεδαστικότητα, δηλ. ίση διασπορά).
- Αυτές οι δύο υποθέσεις είναι βασικές και χωρίς αυτές δεν ισχύουν τα συμπεράσματα του μοντέλου
- Μια επίσης βασική υπόθεση είναι ότι $\mathsf{Cov}(\varepsilon_i, \varepsilon_j) = 0$ δηλαδή το ένα σφάλμα δεν επηρεάζει το άλλο.
- Υπό αυτές τις υποθέσεις έχουμε ότι

$$\mathbb{E}(Y_i) = a + bX_i,\ \mathbb{V}(Y_i) = \sigma^2,\ \mathsf{Cov}(Y_i, Y_j) = 0.$$

- Δηλαδή η γραμμή παλινδρόμησης μας δίνει την αναμενόμενη τιμή της Y για κάθε X.
- Η παράμετρος a είναι η αναμενόμενη τιμή της Y για $X = 0$
- Η παράμετρος b είναι η κλίση της ευθείας της παλινδρόμησης και μας δείχνει τη μεταβολή του Y όταν το X αυξηθεί κατά μια μονάδα.
- Σημείωση: Στο απλό (δηλ. με μία μόνο μεταβλητή) γραμμικό μοντέλο, η γραμμικότητα νοείται ως προς τις παραμέτρους, π.χ. το μοντέλο $Y_i = \beta_1^{\beta_2} X_i$ δεν είναι γραμμικό ενώ το $Y_i = \beta_0 + \beta_1 X_i^2$ είναι.

Ανάλυση παλινδρόμησης: ιδιότητες εκτιμητών.

Θεώρημα 1

Οι εκτιμητές $\hat{a}, \hat{b}$ είναι γραμμικοί συνδυασμοί των Y_i.

Απόδειξη:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i - \sum_{i=1}^{n}(X_i-\bar{X})\bar{Y} \\ &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i - \bar{Y}\underbrace{\sum_{i=1}^{n}(X_i-\bar{X})}_{=0 \text{ γιατί;}} \\ &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i = \sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}_i, \qquad (8)\end{aligned}$$

οπότε προκύπτει άμεσα ότι

$$\hat{b} = \sum_{i=1}^{n} k_i Y_i,\ k_i = \frac{X_i-\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

Ανάλυση παλινδρόμησης: ιδιότητες των k_i.

$$\sum_{i=1}^{n} k_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0 \tag{9}$$

$$\sum_{i=1}^{n} k_i X_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \overset{(*)}{=} 1 \tag{10}$$

$$\sum_{i=1}^{n} k_i^2 = \sum_{i=1}^{n}\left(\frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2 = \frac{\cancel{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{\cancel{2}}}$$

$$= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{11}$$

Όπου στην $(*)$ χρησιμοποιήσαμε ότι

$$(X_i - \bar{X})X_i = X_i^2 - X_i\bar{X} = X_i^2 - X_i\bar{X} + \bar{X}^2 - \bar{X}^2 - X_i\bar{X} + X_i\bar{X} = (X_i - \bar{X})^2 + \bar{X}(X_i - \bar{X}),$$

μαζί με το ότι $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. Επίσης, για το $\hat{a}$ άμεσα συνεπάγεται ότι

Ανάλυση παλινδρόμησης: ιδιότητες εκτιμητών.

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{X}\sum_{i=1}^{n} k_i Y_i = \sum_{i=1}^{n}(\frac{1}{n} - \bar{X}k_i)Y_i.$$

Θεώρημα Gauss-Markov

Για το απλό γραμμικό μοντέλο οι εκτιμητές ελαχίστων τετραγώνων $\hat{a}, \hat{b}$ είναι αμερόληπτοι εκτιμητές των a, b και έχουν ελάχιστη διασπορά μεταξύ όλων των αμερόληπτων εκτιμητών που είναι γραμμικές συναρτήσεις των Y_i

Απόδειξη: Χρησιμοποιούμε την (8) για να γράψουμε

$$\mathbb{E}\hat{b} = \sum_{i=1}^{n} k_i \mathbb{E}Y_i = \sum_{i=1}^{n} k_i(a + bX_i) \overset{(9)-(10)}{=} b \tag{12}$$

$$\mathbb{E}\hat{a} = \mathbb{E}(\bar{Y} - \hat{b}\bar{X}) = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{b}\sum_{i=1}^{n} X_i\right\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Y_i - \mathbb{E}\hat{b}\sum_{i=1}^{n} X_i$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{n}(a + bX_i) - \sum_{i=1}^{n} bX_i\right\} = a \tag{13}$$

Ανάλυση παλινδρόμησης: ιδιότητες εκτιμητών.

Για την διακύμανση των a, b έχουμε

$$\mathbb{V}(\hat{b}) = \sum_{i=1}^{n} k_i^2 \mathbb{V}(Y_i) \overset{(11)}{=} \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{14}$$

$$\begin{aligned}\mathbb{V}(\hat{a}) &= \sum_{i=1}^{n}(\frac{1}{n} - \bar{X}k_i)^2 \mathbb{V}(Y_i) = \sum_{i=1}^{n}\left(\frac{1}{n^2} + (\bar{X}k_i)^2 - \frac{2\bar{X}k_i}{n}\right)\sigma^2 \\ &= \frac{\sigma^2}{n} + \bar{X}^2\sigma^2 \sum_{i=1}^{n} k_i^2 - \frac{2\bar{X}\sigma^2}{n}\sum_{i=1}^{n} k_i \\ &\overset{(9),(11)}{=} \frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\end{aligned} \tag{15}$$

Για να δείξουμε ότι αυτή η διασπορά είναι ελάχιστη μεταξύ όλων των αμερόληπτων εκτιμητών που είναι γραμμικές συναρτήσεις, έστω

$$\hat{b}_1 = \sum_{i=1}^{n} c_i Y_i, \tag{16}$$

η συναρτησιακή μορφή των αμερόληπτων εκτιμητών που είναι γραμμικές συναρτήσεις.

Ανάλυση παλινδρόμησης: ιδιότητες εκτιμητών.

Τότε,

$$\mathbb{V}(\hat{b}_1) = \mathbb{V}(\sum_{i=1}^{n} c_i Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2$$

γιατί αφού έχουμε γραμμικούς εκτιμητές η διαφορά του c_i από το k_i θα είναι μία σταθερά. Έχουμε,

$$\mathbb{V}(\hat{b}_1) = \sigma^2 \sum_{i=1}^{n} (k_i^2 + d_i^2 + 2k_i d_i) \overset{(*)}{=} \mathbb{V}(\hat{b}) + \sigma^2 \sum_{i=1}^{n} d_i^2$$

η οποία ποσότητα ελαχιστοποιείται όταν

$$\sum_{i=1}^{n} d_i^2 = 0 \Leftrightarrow d_i = 0 \ \forall \ i \Leftrightarrow k_i = c_i$$

άρα η εκτιμήτρια ελαχίστων τετραγώνων έχει την ελάχιστη διασπορά μεταξύ όλων των γραμμικών εκτιμητών.

$$(*) : \sum_{i=1}^{n} k_i d_i = \frac{\sum_{i=1}^{n} X_i d_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} - \bar{X} \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = 0$$

Ανάλυση παλινδρόμησης: εκτίμηση του σ^2.

- Ως γνωστόν αν $Y_1, \ldots, Y_n$ τυχαίο δείγμα από κατανομή με μέση τιμή μ και διασπορά σ^2 τότε η εκτιμήτρια του σ^2 είναι η

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$$

- Για μ άγνωστο, χρησιμοποιούμε το $\hat{Y}$ στη θέση του οπότε τώρα

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{17}$$

και ο λόγος που τώρα διαιρούμε με το $n-1$ είναι ότι έχουμε χρησιμοποιήσει το δείγμα για να εκτιμήσουμε το μ οπότε πρέπει να αφαιρέσουμε ένα βαθμό ελευθερίας ώστε η εκτίμηση να είναι αμερόληπτη.

- Στο απλό γραμμικό μοντέλο έχουμε ότι

$$Y_i = a + bX_i + \varepsilon_i, i = 1, \ldots, n \Rightarrow Y_i \sim N(a + bX_i, \sigma^2). \tag{18}$$

- Παρατηρούμε ότι τα Y_i δεν είναι ισόνομα.

Ανάλυση παλινδρόμησης: εκτίμηση του σ^2.

- Ναι μεν όλα ακολουθούν κανονική κατανομή αλλά με διαφορετική μέση τιμή.
- Επίσης, η μέση τιμή του κάθε Y_i είναι το σημείο $a + bX_i$ της ευθείας που προσαρμόσαμε στα δεδομένα.
- Εφαρμόζοντας την (17) στην περίπτωση μας,

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2 \tag{19}$$

όπου διαιρέσαμε με $n-2$ βαθμούς ελευθερίας γιατί έχουμε εκτιμήσει ήδη δύο παραμέτρους.

- Η (19) είναι στην ουσία το μέσο τετραγωνικό σφάλμα του απλού γραμμικού μοντέλου το οποίο ορίζεται από την

$$n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2$$

δηλαδή είναι ο μέσος όρος των υπολοίπων $\hat{\varepsilon}_i$

Ανάλυση παλινδρόμησης: κατανομή των $\hat{b}, \hat{a}$.

- Από την (16) και τη (18) έχουμε ότι ο εκτιμητής $\hat{b}$ είναι γραμμικός συνδυασμός κανονικών τυχαίων μεταβλητών.
- Άρα και η κατανομή του $\hat{b}$ θα είναι κανονική.
- Οι παράμετροι τις κατανομής του $\hat{b}$ είναι φυσικά η μέση τιμή και η διακύμανση που έχουν προσδιοριστεί ήδη στις (12) και (14) οπότε

$$\hat{b} \sim N\left(b, \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right). \tag{20}$$

- Όπου στην (20) εκτιμήσαμε την άγνωστη διασπορά σ^2 με την (19).
- Η κατανομή του $\hat{a} = \bar{Y} - \hat{b}\bar{X}$, είναι πάλι κανονική αφού το $\hat{a}$ είναι γραμμικός συνδυασμός κανονικών τ.μ.
- Από τις (13) και (15) έχουμε ότι

$$\hat{a} \sim N\left(a, \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\right). \tag{21}$$

όπου και πάλι η άγνωστη διασπορά σ^2 εκτιμήθηκε από την (19).

Ανάλυση παλινδρόμησης: κατανομή των $\hat{b}, \hat{a}$.

Πρόταση 1:

Η τ.μ. $(\hat{b} - b)s(\hat{b})^{-1} \sim t_{n-2}$

Απόδειξη: Η απόδειξη βασίζεται στο ότι για δυο ανεξάρτητες τ.μ. Z, U με $Z \sim N(0,1)$ και $U \sim \chi^2_r$ ισχύει ότι $T = Z(U/\tau)^{-1/2} \sim t_r$.
Ξέρουμε ότι όταν $\sigma^2(\hat{b})$ γνωστό, $(\hat{b} - b)\sigma(\hat{b})^{-1} \sim N(0,1)$.
Επίσης,

$$\sigma^{-2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sigma^{-2} \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2 \sim \chi^2_{n-2}.$$

Εφόσον

$$\hat{\sigma}^2 = (n-2)^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \Rightarrow \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

όμως

$$(\hat{b} - b)s(\hat{b})^{-1} = \frac{\frac{\hat{b}-b}{\sigma(\hat{b})}}{\sqrt{\frac{s^2(\hat{b})}{\sigma^2(\hat{b})}}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-2}}} = t_{n-2}$$

Ανάλυση παλινδρόμησης: Συμπερασματολογία για το $\hat{b}$.

Πρόταση:

Να βρεθεί ένα διάστημα εμπιστοσύνης με συντελεστή $1-\alpha$ για τα a, b

Απόδειξη: Από τον ορισμό των δ/των εμπιστοσύνης θέλουμε,

$$P\left(-t_{n-2,1-\alpha/2} \leq (\hat{b}-b)s(\hat{b})^{-1} \leq t_{n-2,1-\alpha/2}\right) = 1-\alpha$$

οπότε το ζητούμενο δ/μα για το b είναι το

$$\left[\hat{b} - s(\hat{b})t_{n-2,1-\alpha/2}, \hat{b} + s(\hat{b})t_{n-2,1-\alpha/2}\right]$$

Παρομοίως για το a έχουμε

$$(\hat{a}-a)s(\hat{a})^{-1} \sim t_{n-2} \tag{22}$$

Οπότε με ακριβώς ίδιο τρόπο, το δ/μα εμπιστοσύνης για το a είναι το

$$[\hat{a} - s(\hat{a})t_{n-2,1-\alpha/2}, \hat{a} + s(\hat{a})t_{n-2,1-\alpha/2}]$$

Ανάλυση παλινδρόμησης: Συμπερασματολογία για το $\hat{b}$.

Πρόταση 2:

Να κατασκευαστεί ένα στατιστικό τεστ επιπέδου $1-\alpha$ για τον έλεγχο της $H_0 : \hat{b} = 0$ και της $H_0 : \hat{a} = 0$

- Από την πρόταση 1, έχουμε ότι $(\hat{b} - b)s(\hat{b})^{-1} \sim t_{n-2}$, δηλαδή το στατιστικό κάτω από την H_0 ακολουθεί κατανομή t με $n-2$ βαθμούς ελευθερίας.
- Ιδιαίτερο ενδιαφέρον έχει η μηδενική υπόθεση $H_0 : b = 0$, δηλαδή ότι η γραμμή παλινδρόμησης είναι οριζόντια και άρα η τ.μ. Y δεν εξαρτάται από την X.
- Απορρίπτουμε την $H_0 : \hat{b} = 0$, έναντι της $H_1 : \hat{b} \neq 0$ για

$$|\hat{b}s^{-1}(\hat{b})| > t_{n-2,\alpha/2}$$

Ομοίως, βάση της (22) έχουμε ότι κρίσιμη περιοχή για τον έλεγχο της $H_0 : \hat{a} = 0$, έναντι της $H_1 : \hat{a} \neq 0$ είναι

$$|\hat{a}s^{-1}(\hat{a})| > t_{n-2,\alpha/2}$$

- Άσκηση: να επεκτείνετε τις H_1 για τις περιπτώσεις $H_1 : \hat{b} \neq b$, $H_1 : \hat{a} \neq a$

Ανάλυση παλινδρόμησης: Παράδειγμα 1 - συνέχεια.

Πρακτική άσκηση 1: συνέχεια

Να υλοποιηθούν οι έλεγχοι για τις εκτιμώμενες παραμέτρους του μοντέλου για το Παράδειγμα 1.

- Μία σύνοψη του εκτιμώμενου μοντέλου δίνεται από την:

```
> summary(al.lm)
Call:
lm(formula = strength ~ alcohol, data = alcoholarm

Residuals:
    Min      1Q  Median      3Q     Max
-8.7847 -2.5450 -0.1477  2.6359  7.4815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.36954    1.20273  21.925  < 2e-16 *
alcohol     -0.29587    0.05105  -5.796 5.14e-07 *
---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “

Residual standard error: 3.874 on 48 degrees of freedom
Multiple R-squared:  0.4117,        Adjusted R-squared:  0.3994
F-statistic: 33.59 on 1 and 48 DF,  p-value: 5.136e-07
```

Η στήλη:
`Pr(>|t|)`
`4.34e-13`
`5.14e-07`
περιέχει τα p-values για τον έλεγχο των $H_0 : \hat{b} = 0$ και $H_0 : \hat{a} = 0$ έναντι των $H_1 : \hat{b} \neq 0$, $H_1 : \hat{a} \neq 0$ αντίστοιχα.

Ιδιαίτερης σημασίας είναι η τιμή του R-squared (δείχνει το % της Y που εξηγείται από την X). Όσο μεγαλύτερο είναι το R^2 τόσο μεγαλύτερη επεξηγηματική ικανότητα έχει το μοντέλο.

Ανάλυση παλινδρόμησης: μετασχηματισμοί γραμμικότητας.

1. Πολλαπλασιαστικό μοντέλο,

$$Y_i = \beta_0 \beta_1^{X_i} \varepsilon_i, \quad \mathbb{E}\varepsilon_i = 1.$$

Θέτουμε $Y_i' = \log_{10} Y_i$ άρα

$$Y_i' = \log_{10}\beta_0 + X_i \log_{10}\beta_1 + \log_{10}\varepsilon_i \Leftrightarrow Y_i' = \gamma_0 + \gamma_1 X_i + \varepsilon_i'$$

όπου θέσαμε $\gamma_j = \log_{10}\beta_j, j = 1, 2$ και για τα τυχαία σφάλματα $\varepsilon_i' = \log_{10}\varepsilon_i$ τώρα έχουμε $\mathbb{E}\varepsilon_i' = 0$.

2. Μοντέλο $Y_i = \beta_0 \beta_1 X_i^{-1} + \varepsilon_i$. Εδώ χρησιμοποιούμε το μετασχηματισμό $X_i' = 1/X_i$ οπότε

$$Y_i' = \beta_0 + \beta_1 X_i' + \varepsilon_i$$

3. Μοντέλο $Y_i = (1 + \exp\{\beta_0 \beta_1 X_i + \varepsilon_i\})^{-1}$. Θέτουμε πρώτα $Y_i' = 1/Y_i$ οπότε

$$Y_i' = 1 + \exp\{\beta_0 \beta_1 X_i + \varepsilon_i\}$$

και συνεχίζουμε θέτοντας $Y_i'' = \ln(Y_i' - 1)$ οπότε τελικά $Y_i'' = \beta_0 + \beta_1 X_i + \varepsilon_i$.

Ανάλυση παλινδρόμησης: ιδιότητες υπολοίπων.

- Τα υπόλοιπα $\hat{\varepsilon}_i = \hat{Y}_i - Y_i, i = 1, \ldots, n$ είναι πολύ χρήσιμα στον έλεγχο καταλληλότητας του μοντέλου.
- Τα υπόλοιπα είναι οι κατακόρυφες αποκλίσεις από την προσαρμοσμένη ευθεία παλινδρόμησης.
- Φυσικά έχουμε $\hat{\varepsilon}_i \neq \varepsilon_i$ γιατί τα τυχαία σφάλματα ε_i εξαρτώνται από τα πραγματικά a, b τα οποία είναι άγνωστα.

Ιδιότητες των υπολοίπων:

1. Πάντα ισχύει: $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$ - αν η ιδιότητα δεν επαληθεύεται τότε η ευθεία που έχουμε εκτιμήσει είναι λάθος.
2. $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ είναι ελάχιστο. Επαληθεύεται αυτόματα αφού η ευθεία εκτιμάται με βάση αυτόν τον περιορισμό.
3. $\sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0$ γιατί, λόγω της (6),

$$\sum_{i=1}^{n} X_i \hat{\varepsilon}_i = \sum_{i=1}^{n} X_i (Y_i - \hat{a} - \hat{b} X_i) = \sum_{i=1}^{n} X_i Y_i - \hat{a} \sum_{i=1}^{n} X_i - \hat{b} \sum_{i=1}^{n} X_i^2 = 0$$

4. $\sum_{i=1}^{n} Y_i \hat{\varepsilon}_i = 0$ - συνέπεια της

Ανάλυση παλινδρόμησης: ιδιότητες υπολοίπων.

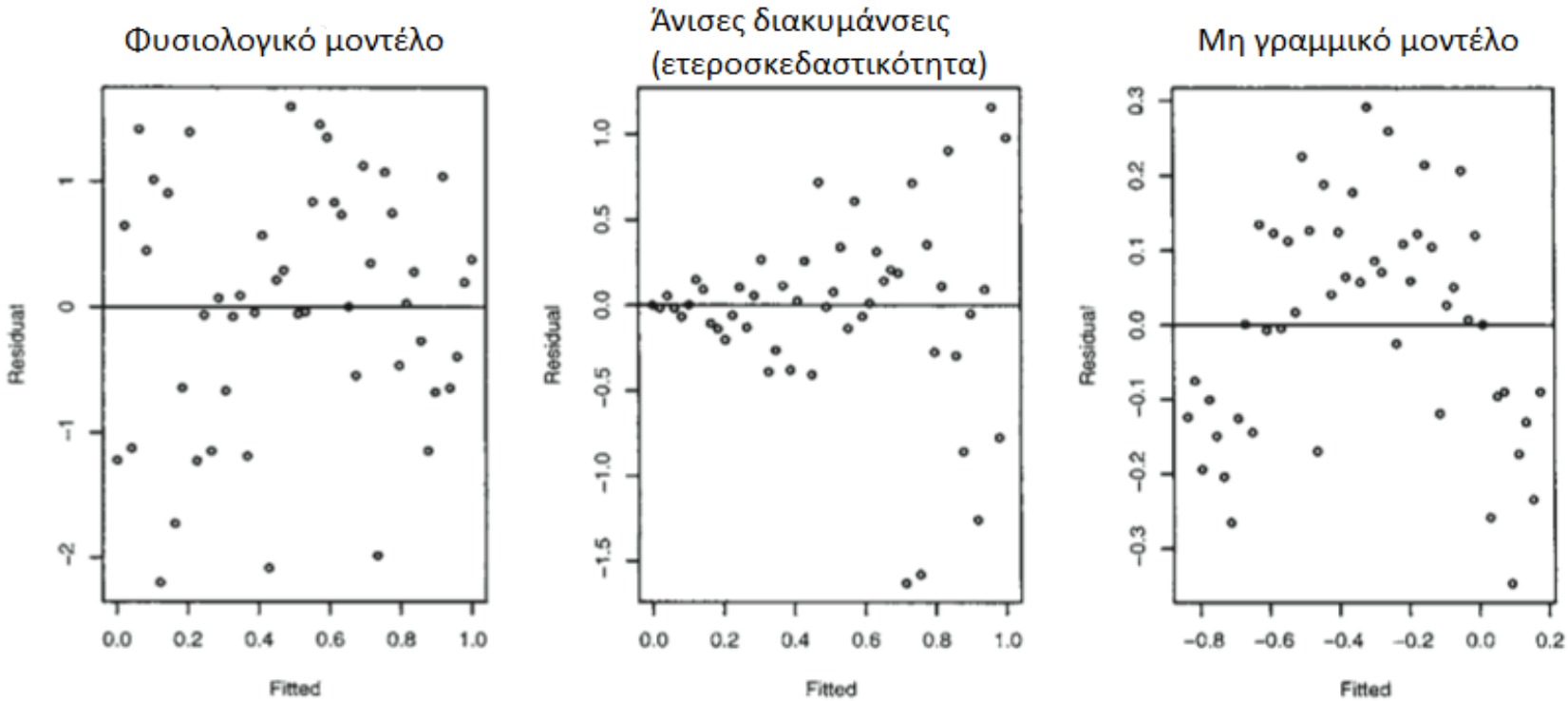
$$\sum_{i=1}^{n}(\hat{a}+\hat{b}X_i)(Y_i-\hat{Y}_i) = \hat{a}\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)}_{=0} + \hat{b}\sum_{i=1}^{n}X_i(Y_i-\hat{Y}_i)$$

$$= \underbrace{\hat{b}\sum_{i=1}^{n}X_iY_i - \hat{a}\sum_{i=1}^{n}\hat{X}_i - \hat{b}\sum_{i=1}^{n}X_i^2}_{\text{λόγω της (6)}} = 0$$

- Τα υπόλοιπα, $\hat{\varepsilon}_i$, στην ουσία είναι εκτιμήσεις των ε_i με μέση τιμή 0, ίση διασπορά σ^2 και συνδιακύμανση 0.
- Με άλλα λόγια τα υπόλοιπα πρέπει να είναι τυχαία κατανεμημένα γύρω από το 0, ασυσχέτιστα και ομοσκεδαστικά.
- Αν κάτι από τα παραπάνω (έστω και ένα) δεν ισχύει, τότε δεν ισχύει και το στατιστικό μοντέλο που εκτιμήσαμε.

Ανάλυση παλινδρόμησης: ιδιότητες υπολοίπων.

- Οι τρεις περιπτώσεις υπολοίπων στην παρακάτω εικόνα δείχνουν τους τύπους υπολοίπων που μπορεί να συναντήσουμε στην πράξη.



- Στην περίπτωση του φυσιολογικού μοντέλου, βλέπουμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρο από τον οριζόντιο άξονα.
- Στην περίπτωση της ετεροσκεδαστικότητας, βλέπουμε ότι η διακύμανση των υπολοίπων αυξάνει όσο αυξάνει το x.
- Στην περίπτωση της μη γραμμικής σχέσης τα υπόλοιπα έχουν μια ξεκάθαρη πατέντα (μη τυχαία).

Ανάλυση παλινδρόμησης: συμπερασματολογία.

- Η παλινδρόμηση εξηγεί ως ένα βαθμό τη μεταβλητότητα του Y
- Όταν η ευθεία της παλινδρόμησης ισχύει, τότε αυτό που λέμε είναι στην ουσία ότι το Y μεταβάλλεται με το X
  - δηλ. μειώνεται όταν αυξάνεται το X (για αρνητικό $\hat{b}$)
  - δηλ. αυξάνεται όταν αυξάνεται το X (για θετικό $\hat{b}$)
- Η παλινδρόμηση δεν συστήνει ότι η X προκαλεί μεταβολές στην Y (μπορεί π.χ. και οι δυο να συνδέονται με μια τρίτη μεταβλητή και οι μεταβολές να προκαλούνται από εκείνη)
- Αν η $H_0 : \hat{b} = 0$ δεν απορρίπτεται τότε έχουμε ένδειξη ότι οι X, Y δεν συνδέονται γραμμικά.
  - αυτό όμως δεν σημαίνει ότι δεν συνδέονται με κάποια άλλη σχέση, π.χ. με μια μη γραμμική, καμπυλόγραμμη μορφή.
- ένας γραφικός τρόπος για έλεγχο της γραμμικότητας του μοντέλου είναι η γραφική παράσταση των X_i έναντι των υπολοίπων ε_i
  - αν τα υπόλοιπα σχηματίζουν μια οριζόντια ζώνη γύρω από το 0 χωρίς κάποια συστηματική τάση, αυτό είναι ένδειξη γραμμικότητας.
  - Επίσης, αν όσο αυξάνεται η X αυξάνεται και η διακύμανση των υπολοίπων αυτό είναι ένδειξη ότι τα σφάλματα δεν έχουν ίσες διακυμάνσεις.