9

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

9.1. Εισαγωγή

9.1.1. Απλή γραμμική παλινδρόμηση

Η γραμμική παλινδρόμηση είναι μία στατιστική τεχνική που μας δίνει τη δυνατότητα να διαπιστώσουμε τον τρόπο με τον οποίο μια μεταβλητή που ονομάζεται ***ανεξάρτητη μεταβλητή (Χ)*** επηρεάζει τις τιμές μιας άλλης μεταβλητής που ονομάζεται ***εξαρτημένη μεταβλητή (Υ)***, και αυτή η μορφή γραμμικής παλινδρόμησης ονομάζεται ***«απλή γραμμική παλινδρόμηση»***. Η γραμμική παλινδρόμηση, λοιπόν, μοιάζει αρκετά με την απλή συσχέτιση που αναφέρεται στο Κεφάλαιο 7. Όμως, ενώ η απλή συσχέτιση μας πληροφορεί μόνο για το αν υπάρχει γραμμική συσχέτιση ανάμεσα στις 2 μεταβλητές (ένταση και διεύθυνση της σχέσης), η γραμμική παλινδρόμηση απαντά και στο ερώτημα «Πόσο πολύ θα μεταβληθεί η Υ όταν θα αλλάξει η Χ». Με άλλα λόγια, με τη γραμμική παλινδρόμηση μπορούμε να εκτιμήσουμε πόσο πολύ θα αλλάξει η Υ για συγκεκριμένη μεταβολή της Χ. Συνεπώς, λοιπόν, η γραμμική παλινδρόμηση μας δίνει τη δυνατότητα να προβλέψουμε τις τιμές της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή παίρνει συγκεκριμένες τιμές. Για να επιτευχθεί αυτό, το μόνο που απαιτείται είναι να εκφραστεί αυτή η σχέση μεταξύ των Χ και Υ με μία κατάλληλη ***μαθηματική συνάρτηση***.

Σημειώση: Μπορούν να χρησιμοποιηθούν και περισσότερες από μία ανεξάρτητες μεταβλητές για να προβλεφθούν οι τιμές της εξαρτημένης μεταβλητής και σε αυτή την περίπτωση η γραμμική παλινδρόμηση ονομάζεται ***«πολλαπλή γραμμική παλινδρόμηση»***.

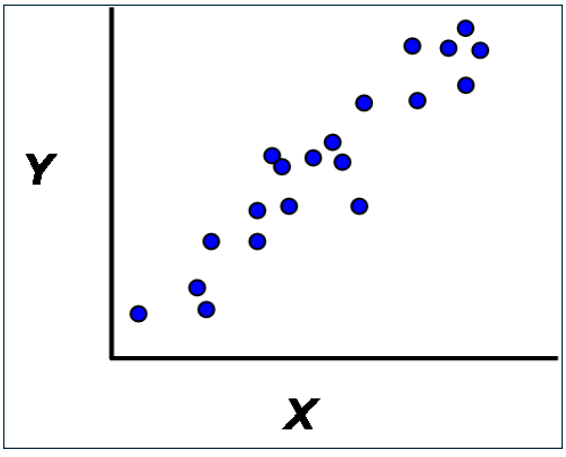
9.1.1.1. Προσαρμογή της απλής γραμμικής παλινδρόμησης

Αυτή η κατάλληλη μαθηματική συνάρτηση είναι της μορφής:

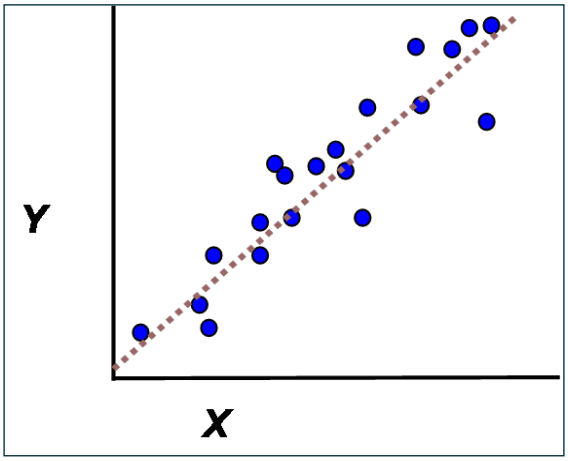
$$\hat{Y} = \beta_0 + \beta_1 X_1 \quad (1)$$

και **γεωμετρικά**, μεταφράζεται ως μία ***«ευθεία γραμμή»*** που θα διέρχεται μέσα από τα πραγματικά δεδομένα των μεταβλητών Χ και Υ. Οι συντελεστές β_0, β_1 είναι άγνωστοι και πρέπει να εκτιμηθούν προκειμένου για κάθε τιμή της Χ να είναι εφικτή ο υπολογισμός της Υ. Είναι γεγονός, ότι οι μαθηματικές συναρτήσεις που μπορούν να υπολογιστούν είναι άπειρες (άπειροι οι συνδυασμοί των β_0, β_1 που μπορούν να χρησιμοποιηθούν), όμως, μόνο μία είναι η βέλτιστη, δηλαδή αυτή που περιγράφει με τον καλύτερο δυνατό τρόπο την πραγματική σχέση ανάμεσα στην Χ και την Υ. Συνεπώς, λοιπόν, οι β_0, β_1 θα πρέπει να εκτιμηθούν με σκοπό η μαθηματική συνάρτηση που θα προκύψει να είναι η βέλτιστη.

Ας υποθέσουμε, λοιπόν, ότι στην **Εικόνα 9.1** παρουσιάζονται τα πραγματικά δεδομένα των μεταβλητών Χ και Υ σε ένα δείγμα. Και ας υποθέσουμε ότι στην **Εικόνα 9.2** παρουσιάζεται η ***«ευθεία γραμμή»*** που προέκυψε από την προσαρμογή μίας τυχαία γραμμικής μαθηματικής συνάρτησης. Για να είναι η γραμμική συνάρτηση της **Εικόνας 9.2** η βέλτιστη, θα πρέπει η απόσταση της ***«ευθείας γραμμής»*** από όλα τα σημεία να είναι η ελάχιστη δυνατή. Λαμβάνοντας υπόψη ότι η απόσταση κάθε σημείου της **Εικόνας 9.2** από την ευθεία γραμμή ονομάζεται ***«σφάλμα πρόβλεψης»***, αντιλαμβανόμαστε ότι η ***«ευθεία γραμμή»*** που περιγράφει καλύτερα τα δεδομένα είναι αυτή που ελαχιστοποιεί τα σφάλματα πρόβλεψης. Συνεπώς, οι συντελεστές β_0, β_1 της γραμμικής συνάρτησης (1) θα πρέπει να εκτιμηθούν με γνώμονα να ελαχιστοποιούνται τα ***«σφάλματα πρόβλεψης»***. Η μέθοδος που χρησιμοποιείται για την εκτίμηση των συντελεστών β_0, β_1 ονομάζεται ***«μέθοδος των ελαχίστων τετραγώνων»***.



Εικόνα 9.1. Στικτόγραμμα των μεταβλητών Χ και Υ.



Εικόνα 9.2. Προσαρμογή της ευθείας γραμμής και το σφάλμα πρόβλεψης από την προσαρμογή αυτής.

9.1.1.2. Ερμηνεία συντελεστών της απλής γραμμικής παλινδρόμησης και ο έλεγχος υποθέσεων για τους συντελεστές

Οι συντελεστές β_0 , β_1 *ερμηνεύονται* ως εξής:

β_0: η αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y, *όταν η τιμή* της μεταβλητής X είναι μηδέν.

β_1: η μεταβολή στην εξαρτημένη μεταβλητή Y, για κάθε μονάδα αύξηση της ανεξάρτητης μεταβλητής X_1.

Επιπλέον, ενδιαφέρον παρουσιάζει ο στατιστικός έλεγχος για τις πραγματικές τιμές των β_0, β_1 στον πληθυσμό. Πιο συγκεκριμένα, μπορούμε να ελέγξουμε αν οι τιμές των συντελεστών β_0, β_1 διαφέρουν από το μηδέν στον πραγματικό πληθυσμό. Ιδιαίτερο ενδιαφέρον, παρουσιάζει ο έλεγχος υποθέσεων για το β_1, από όπου θα προκύψει και το συμπέρασμα αν η ανεξάρτητη μεταβλητή X συνεισφέρει σημαντικά στην πρόβλεψη των τιμών της μεταβλητής Y. Συνεπώς, λοιπόν, οι μηδενικές και εναλλακτικές υποθέσεις για τους συντελεστές β_0, β_1 είναι:

H_0: $\beta_0 = 0$ και H_1: $\beta_0 \neq 0$

H_0: $\beta_1 = 0$ και H_1: $\beta_1 \neq 0$

Για να καταλήξουμε στο συμπέρασμα ότι διαφέρουν τα β_0, β_1 από το μηδέν, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_0 (δηλαδή το p-value) να είναι $<\alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Αν δεν απορρίψουμε την H_0 από τον έλεγχο υποθέσεων για το β_1 θα συμβαίνει ένα από τα 2:

α. η μεταβλητή X θα είναι ελάχιστα ή καθόλου σημαντική για την πρόβλεψη της Y.

β. η πραγματική σχέση ανάμεσα στη X και την Y δεν είναι γραμμική.

9.1.2. Πολλαπλή γραμμική παλινδρόμηση

Όπως έχει ήδη αναφερθεί στη σημείωση της παραγράφου 9.1.1, η πολλαπλή παλινδρόμηση είναι η επέκταση της απλής παλινδρόμησης στην περίπτωση που έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές. Η εισαγωγή περισσότερων ερμηνευτικών μεταβλητών έχει ως σκοπό την ερμηνεία όλο και μεγαλύτερου τμήματος της συνολικής μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής Y, μειώνοντας κατ' αυτόν τον τρόπο τις τιμές των σφαλμάτων e_i, άρα και τη διακύμανσή τους $\hat{\sigma}^2$.

Με την πολλαπλή γραμμική παλινδρόμηση καθίστανται εφικτά τα εξής:

- Εκτίμηση της επίδρασης της μεταβολής κάποιας ανεξάρτητης (ερμηνευτικής) μεταβλητής στην εξαρτημένη μεταβλητή Y, ελέγχοντας για την ενδεχόμενη επίδραση άλλων μεταβλητών, δηλαδή εκτίμηση της απευθείας επίδρασης μιας μεταβλητής στην τιμή της Y.
- Πιο ακριβής πρόβλεψη της τιμής της εξαρτημένης μεταβλητής για κάποια μελλοντική παρατήρηση.

Στην απλή γραμμική παλινδρόμηση προσαρμόζουμε στο δείγμα των παρατηρήσεων (X_i, Y_i), i=1,2,...,n την καλύτερη ευθεία. Αν έχουμε δύο ανεξάρτητες μεταβλητές, τότε οι παρατηρήσεις μας θα είναι διατεταγμένες τριάδες (Y_i, X_{i1}, X_{i2}), i = 1,2,...,n και θα αντιπροσωπεύουν σημεία στον χώρο των τριών διαστάσεων. Στα σημεία αυτά θα προσαρμόσουμε το *«επίπεδο ελάχιστων τετραγώνων»*. Γενικά, αν έχουμε k ανεξάρτητες μεταβλητές, τότε οι παρατηρήσεις (Y_i, X_{i1}, X_{i2},..., X_{ik}), i = 1,2,...,n , θα είναι σημεία του χώρου των k+1 διαστάσεων και στα σημεία αυτά θα προσαρμόσουμε το *«πολυεπίπεδο ελάχιστων τετραγώνων»*.

9.1.2.1. Εκτίμηση των παραμέτρων του υποδείγματος της πολλαπλής γραμμικής παλινδρόμησης

Στην πολλαπλή γραμμική παλινδρόμηση, όπως και στην απλή, θεωρούμε ότι η εξαρτημένη μεταβλητή Y μπορεί να εκφραστεί ως μια γραμμική συνάρτηση των k ανεξάρτητων μεταβλητών X_1, X_2,..., X_k.

Το μοντέλο της γραμμικής παλινδρόμησης επιδιώκει τον εντοπισμό της κατάλληλης γραμμικής σχέσης μεταξύ της εξαρτημένης μεταβλητής Y και των k ανεξάρτητων μεταβλητών X_1, X_2,..., X_k. Ας υποθέσουμε ότι η γραμμική σχέση που συνδέει τις παραπάνω μεταβλητές, εκφράζεται από την παρακάτω συνάρτηση:

$$\hat{Y}_i \equiv \mu_{Y\backslash X_1, X_2, \ldots, X_k} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \quad (2)$$

όπου

- Y_i: η τιμή της εξαρτημένης μεταβλητής (για την οποία ενδιαφερόμαστε να διατυπώσουμε προβλέψεις) στην i παρατήρηση
- X_{ij}: η τιμή της j ανεξάρτητης μεταβλητής για την i παρατήρηση, i = 1, 2,..., n & j = 1, 2, k
- β_0: σταθερά, που εκφράζει τη μέση τιμή της εξαρτημένης μεταβλητής όταν όλες οι ερμηνευτικές μεταβλητές έχουν τιμή μηδέν.
- $\beta_1, \beta_2, \ldots, \beta_k$: οι μερικοί συντελεστές παλινδρόμησης για τις μεταβλητές $X_1, \ldots, X_k$, αντίστοιχα, κάθε ένας από τους οποίους εκφράζει την κατά μέσο όρο μεταβολή της τιμής της εξαρτημένης μεταβλητής όταν η αντίστοιχη μεταβλητή μεταβάλλεται κατά μία μονάδα και οι υπόλοιπες παραμένουν σταθερές.

9.1.2.2. Έλεγχος υποθέσεων στην πολλαπλή γραμμική παλινδρόμηση

Σε ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να ανακύψουν τα εξής ερωτήματα:

- Κατά πόσο ολόκληρο το μοντέλο συνεισφέρει στατιστικά σημαντικά στην πρόβλεψη της εξαρτημένης μεταβλητής Y ή αλλιώς στην ερμηνεία της μεταβλητότητάς της.
- Κατά πόσο μια συγκεκριμένη τυχαία μεταβλητή παρέχει στατιστικά σημαντική πληροφορία για την πρόβλεψη της Y, δεδομένης της παρουσίας άλλων ερμηνευτικών μεταβλητών στο μοντέλο.
- Κατά πόσο μια ομάδα τυχαίων μεταβλητών παρέχει στατιστικά σημαντική πληροφορία για την πρόβλεψη της Y, δεδομένης της παρουσίας άλλων ερμηνευτικών μεταβλητών στο μοντέλο.

Η απάντηση σε όλα τα παραπάνω ερωτήματα, δίνεται πραγματοποιώντας τους κατάλληλους στατιστικούς ελέγχους υποθέσεων.

Α. Έλεγχος για ολόκληρο το μοντέλο

Η διατύπωση των κατάλληλων υποθέσεων για την πραγματοποίηση αυτού του ελέγχου είναι η εξής:

H_0: Όλες οι μεταβλητές που συμμετέχουν στο μοντέλο δεν ερμηνεύουν στατιστικά σημαντικό μέρος της μεταβλητότητας των δεδομένων, δηλαδή δεν συμβάλλουν στην πρόβλεψη της εξαρτημένης μεταβλητής, ή αλλιώς

$$\beta_1 = \beta_2 = \ldots = \beta_k = 0$$

H_1: Έστω και μια μεταβλητή ερμηνεύει στατιστικά σημαντικό μέρος της μετα-

βλητότητας των δεδομένων, δηλαδή έστω και ένας από τους συντελεστές β_1, β_2,..., β_k είναι διάφορος του μηδενός.

Β. Έλεγχος για την προσθήκη μίας μόνο μεταβλητής

Έστω ότι σε ένα αρχικό μοντέλο που περιέχει k ερμηνευτικές μεταβλητές προσθέτουμε άλλη μία (X_{k+1}). Αυτό που μας ενδιαφέρει είναι να ελέγξουμε αν αυτή η μεταβλητή προσφέρει στατιστικά σημαντική πληροφορία στην πρόβλεψη της εξαρτημένης μεταβλητής Y, δεδομένης της παρουσίας των άλλων ερμηνευτικών μεταβλητών.

Άρα, ο κατάλληλος έλεγχος υποθέσεων διατυπώνεται ως εξής:

H_0: Η προσθήκη της μεταβλητής X_{k+1} στο μοντέλο, δεδομένης της παρουσίας των υπολοίπων k μεταβλητών, δε βελτιώνει στατιστικά σημαντικά την πρόβλεψη της εξαρτημένης μεταβλητής, δηλαδή δεν αυξάνει στατιστικά σημαντικά την ερμηνευτική ικανότητα του μοντέλου παλινδρόμησης, ή αλλιώς: $\beta_{k+1} = 0$

H_1: Η προσθήκη της μεταβλητής X_{k+1} στο μοντέλο βελτιώνει στατιστικά σημαντικά την πρόβλεψη της εξαρτημένης μεταβλητής, ή αλλιώς: $\beta_{k+1} \neq 0$

Γ. Έλεγχος για την προσθήκη μίας ομάδας μεταβλητών

Έστω ότι σε ένα μοντέλο παλινδρόμησης που περιέχει k ερμηνευτικές μεταβλητές, προσθέτουμε άλλες m μεταβλητές και θέλουμε να ελέγξουμε αν αυτές οι επιπλέον m μεταβλητές βελτιώνουν την ερμηνευτική-προβλεπτική ικανότητα του μοντέλου.

Ο κατάλληλος έλεγχος υποθέσεων διατυπώνεται ως εξής:

H_0: Και οι m μεταβλητές, συνολικά, δεν βελτιώνουν στατιστικά σημαντικά την ερμηνευτική-προβλεπτική ικανότητα του μοντέλου, ή αλλιώς

$$\beta_{k+1} = \beta_{k+2} = \ldots = \beta_m = 0$$

H_1: Τουλάχιστον μία από τις m μεταβλητές βελτιώνει στατιστικά σημαντικά την προβλεπτική ικανότητα του μοντέλου, ή αλλιώς ένας από τους παραπάνω συντελεστές (β_{k+1}, β, ..., β_m) είναι διάφορος του μηδενός.

9.1.3. Προϋποθέσεις ορθής εφαρμογής της γραμμικής παλινδρόμησης

Προκειμένου να εφαρμοστεί ορθά η γραμμική παλινδρόμηση (απλή η πολλαπλή) θα πρέπει να ισχύουν οι εξής προϋποθέσεις:

i. *Γραμμικότητα:* Η συσχέτιση ανάμεσα σε κάθε μία από τις ανεξάρτητες μεταβλητές και την εξαρτημένη μεταβλητή θα πρέπει να είναι γραμμική.

ii. *Κανονικότητα:* Η κατανομή των σφαλμάτων να είναι κανονική για κάθε τιμή των ανεξάρτητων μεταβλητών.

iii. *Ομοσκεδαστικότητα:* Η τυπική απόκλιση των σφαλμάτων να είναι ίση για όλες τις τιμές της κάθε ανεξάρτητης μεταβλητής.

iv. *Ανεξαρτησία:* Οι παρατηρήσεις θα πρέπει να είναι ανεξάρτητες, δηλαδή να προέρχονται από διαφορετικά άτομα.

v. *Πολυσυγγραμμικότητα:* Οι ανεξάρτητες μεταβλητές δεν πρέπει να συσχετίζονται ισχυρά μεταξύ τους. Η συσχέτιση μεταξύ των μεταβλητών που χρησιμοποιούνται σε πολλαπλή γραμμική παλινδρόμηση ως ανεξάρτητες θα πρέπει να είναι η μικρότερη δυνατή, δεδομένου ότι ισχυρή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών δημιουργεί το πρόβλημα της πολυσυγγραμμικότητας το οποίο με τη σειρά του έχει ως αποτέλεσμα τον υπολογισμό εκτιμητών με αυξημένα τυπικά σφάλματα. Συνεπώς, λοιπόν, είναι απαραίτητο να ελέγχουμε τον βαθμό συσχέτισης των ανεξάρτητων μεταβλητών που χρησιμοποιούνται σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης. Τα κατάλληλα στατιστικά γι' αυτόν τον έλεγχο είναι το ***Tolerance*** και το ***VIF***.

   - ***Tolerance:*** όσο μεγαλύτερο είναι τόσο μικρότερη είναι η συσχέτισή του με όλες τις υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου. Το εύρος τιμών του ***Tolerance*** διακυμαίνεται μεταξύ 0 και 1. Τιμές πολύ κοντά στην 1 υποδηλώνουν την έλλειψη συσχέτισης ανάμεσα στις ανεξάρτητες μεταβλητές.

   - ***VIF:*** όσο μεγαλύτερο είναι τόσο μεγαλύτερη είναι η συσχέτιση του παράγοντα με τις υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου. Το εύρος τιμών του VIF είναι από 1 έως +άπειρο. Τιμές πάνω από 2 ή 3 υποδηλώνουν ισχυρή συσχέτιση.

Οι προϋποθέσεις (i) έως (iv), μπορούν να ελεγχθούν με τους εξής τρόπους:

i. Παριστάνοντας γραφικά τα κατάλοιπα έναντι των εκτιμούμενων τιμών που έχουν προκύψει βάσει του μοντέλου, για να ελέγξουμε αν υπάρχει κάποια καμπύλη στο γράφημα και για να δούμε αν τα κατάλοιπα βρίσκονται γύρω από το μηδέν και έχουν ίση διακύμανση κατά μήκος όλων των εκτιμούμενων τιμών.

ii. Φτιάχνοντας ιστόγραμμα ή γράφημα κανονικής πιθανότητας των καταλοίπων. Το ιστόγραμμα, αν ισχύει η προϋπόθεση της κανονικότητας θα πρέπει να είναι συμμετρικό, ενώ στο γράφημα κανονικής πιθανότητας, τα κατάλοιπα θα πρέπει να βρίσκονται πάνω σε μία ευθεία διαγώνια γραμμή.

9.1.4. Ερμηνευτικότητα του μοντέλου

Τόσο στην απλή όσο και στην πολλαπλή γραμμική παλινδρόμηση, ο συντελεστής προσδιορισμού (R^2) εκφράζει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητή Y που ερμηνεύεται από το μοντέλο ή αλλιώς από την μία ανεξάρτητη μεταβλητή (στην απλή γραμμική παλινδρόμηση) ή όλες τις ανεξάρτητες μεταβλητές (στην πολλαπλή γραμμική παλινδρόμηση). Το εύρος τιμών του R^2 είναι: