ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΤΟΥΡΙΣΜΟΥ

ΠΠΣ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ ΜΕΣΟΛΟΓΓΙ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Αριστείδης Καραμάνος

Επιβλέπουσα Καθηγήτρια

Βασιλική Καρυώτη

Μεσολόγγι 2022

UNIVERSITY OF PATRAS

SCHOOL OF ECONOMICS & BUSINESS ADMINISTRATION

DEPARTMENT OF TOURISM ADMINISTRATION

FORMER DEPARTMENT OF ACCOUNTING AND FINANCE
MESSOLONGHI

THESIS

REGRESSION MODELS

Aristides Karamanos

Messolonghi 2022

Οι διαπιστώσεις, τα αποτελέσματα, τα συμπεράσματα και οι πιθανές προτάσεις της παρούσας Πτυχιακής Εργασίας, εκτός των αναφορών που σημαίνονται ως λήμματα, αποτελούν προσωπικές θεωρητικές ή εμπειρικές διαπιστώσεις του φοιτητή/φοιτήτριας ή της ομάδας των φοιτητών που την επιμελήθηκαν και δεν απηχούν κατ' ανάγκη τη γνώμη του εισηγητή εκπαιδευτικού, ή του Εκπαιδευτικού Προσωπικού του Τμήματος Διοίκησης Τουρισμού (του πρώην Λογιστικής & Χρηματοοικονομικής του Α.Τ.Ε.Ι. Δυτ. Ελλάδας) του Πανεπιστήμιου Πατρών.

ΕΥΧΑΡΙΣΤΙΕΣ

Με αφορμή την ολοκλήρωση της παρούσας πτυχιακής εργασίας θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια μου, κα. Βασιλική Καρυώτη για τη βοήθειά της στην ολοκλήρωση της πτυχιακής εργασίας μου. Οφείλω το μεγαλύτερο ευχαριστώ στην οικογένειά μου για την στήριξη και υπομονή σε όλη τη διάρκεια των σπουδών μου. Τέλος, θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν στην μέχρι τώρα εκπαίδευσή μου.

ΠΕΡΙΛΗΨΗ

Είναι κοινά αποδεκτό πως το στατιστικό μοντέλο αποτελεί μια τυποποίηση στοχαστικών σχέσεων μεταξύ μεταβλητών, όπως αυτές παρουσιάζονται με τη μορφή μαθηματικών σχέσεων και εξισώσεων. Ο απώτερος στόχος από μια στατιστική ανάλυση είναι να επιτυγχάνεται η πιο ακριβή περιγραφή ενός συστήματος (φαινομένου ή γεγονότος). Σχεδόν σε κάθε φυσικό ή τεχνητό σύστημα, υπάρχουν μεταβλητές των οποίων οι ποσότητες συνεχώς αλλάζουν. Ανέκαθεν εύλογο ήταν εκείνο το ερώτημα για την μελέτη του βαθμού επίδρασης που οι περισσότερες ενεργούν και με ποιον τρόπο πάνω σε άλλες. Η μελέτη αυτή είναι το αντικείμενο της ανάλυσης παλινδρόμησης, μίας ευρέως χρησιμοποιούμενης στατιστικής τεχνικής, η οποία χρησιμοποιείται για την δημιουργία των κατάλληλων μοντέλων σχέσεων και εξαρτήσεων μεταξύ μεταβλητών. Τα διάφορα στατιστικά μοντέλα παλινδρόμησης, βασίζονται σε κάποιες βασικές υποθέσεις, τις οποίες οι ερευνητές υποχρεούνται να ελέγχουν πριν την ανάλυση του μοντέλου, δημιουργώντας τις κατάλληλες συνθήκες για την επίλυση τους. Εντούτοις, οι υποθέσεις αυτές συχνά παραβιάζονται και ειδικότερα όταν τα δεδομένα συλλέγονται από τον πραγματικό κόσμο ενώ υπάρχουν και εξωγενείς παράγοντες που δεν μπορούν να συνεκτιμηθούν.

Στην παρούσα πτυχιακή εργασία γίνεται μια συνολική προσπάθεια αποτύπωσης των τεχνικών ανάλυσης της παλινδρόμησης, ώστε να γίνουν κατανοητές οι διάφορες μορφές των μοντέλων παλινδρόμησης. Προκειμένου να επιτευχθεί αυτό, παρουσιάζονται παράλληλα και κατάλληλα παραδείγματα που βοηθούν στην κατανόηση μαθηματικών εννοιών. Είναι σημαντικό να αναφερθεί πως για κάποιον που αναζητά να αποκτήσει μια ολοκληρωμένη εικόνα για τα στατιστικά μοντέλα, είναι αναγκαίο να τα εξετάζει πάντα στον πραγματικό κόσμο για να έχει την δυνατότητα να μειώνει τις τυχόν παραβιάσεις των κανόνων στο μοντέλο, και παράλληλα να προβλέπει το βαθμό στατιστικής σημαντικότητας.

ABSTRACT

It is commonly accepted that the statistical model is a standardization of stochastic relations between variables, as they are presented in the form of mathematical relations and equations. The ultimate goal of a statistical analysis is to achieve the most accurate description of a system (phenomenon or event). In almost every natural or artificial system, there are variables whose quantities are constantly changing. It has always been a reasonable question to study the degree to which most people act and how they act on others. This study is the subject of regression analysis, a widely used statistical technique used to generate appropriate models of relationships and dependencies between variables. The various statistical regression models are based on some basic assumptions, which researchers are required to check before analyzing the model, creating the appropriate conditions for their solution. However, these assumptions are often violated, especially when the data is collected from the real world, and there are external factors that cannot be taken into account.

In the present dissertation, a comprehensive attempt is made to capture the regression analysis techniques, in order to understand the various forms of regression models. In order to achieve this, appropriate examples are presented that help to understand mathematical concepts. It is important to note that for someone looking to get a complete picture of statistical models, it is necessary to always look at them in the real world in order to be able to reduce any violations of the rules in the model, while predicting the degree of statistical significance.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΙΣΑΓΩΓΗ

Αντικείμενο της παρούσας πτυχιακής εργασίας αποτελεί η ανάλυση των μοντέλων παλινδρόμησης. Η ανάλυση παλινδρόμησης επιχειρεί να συλλάβει και να αποτυπώσει συσχετίσεις μεταξύ δεδομένων με στόχο την εξήγηση και την πρόβλεψη των τιμών των μεταβλητών, όταν ισχύουν κάποιες συγκεκριμένες συνθήκες.

Αξίζει να σημειωθεί ότι η παλινδρόμηση, ως μια στατιστική τεχνική μοντελοποίησης έχει ως στόχο τον ποσοτικό προσδιορισμό των σχέσεων μεταξύ μεταβλητών, αποτελεί το μέσο για την εξαγωγή χρήσιμων συμπερασμάτων και τη λήψη ορθών αποφάσεων, η οποία είναι απαραίτητη σε όλους τους τομείς της ανθρώπινης δραστηριότητας. Η σπουδαιότητα του εν λόγω θέματος, λοιπόν, έγκειται στο γεγονός ότι οι μέθοδοι προσδιορισμού και αξιολόγησης των σχέσεων μεταξύ τέτοιων μεταβλητών χρησιμοποιούνται ευρέως σε πολλούς διαφορετικούς χώρους, όπως τα οικονομικά, το μάρκετινγκ (marketing), την ιατρική και τις θετικές επιστήμες.

Στόχος της εργασίας είναι η ορθή και όσο το δυνατόν πληρέστερη ανάλυση του θέματος, ώστε με συνοπτικό και εύληπτο τρόπο να γίνουν αντιληπτές οι διάφορες μορφές των μοντέλων παλινδρόμησης. Για το λόγο αυτό θα παραθέσω διαγραμματικές παραστάσεις, γραφήματα και πληθώρα παραδειγμάτων που θα διευκολύνουν την κατανόηση μαθηματικών εννοιών.

Η όλη δομή της εργασίας έχει ως ακολούθως.

Περιλαμβάνει πέντε κεφάλαια, στο πρώτο εκ των οποίων αναλύεται η απλή γραμμική παλινδρόμηση, η οποία αποτελεί ένα από τα πιο διαδεδομένα μοντέλα παλινδρόμησης.

Το δεύτερο κεφάλαιο επικεντρώνεται σ’ ένα άλλο μοντέλο παλινδρόμησης, την πολλαπλή γραμμική παλινδρόμηση.

Στο τρίτο και τέταρτο κεφάλαιο αναφέρεται το απλό λογιστικό μοντέλο και η πολλαπλή λογιστική παλινδρόμηση αντίστοιχα, τα οποία θα προσπαθήσω να αναλύσω με εύληπτο και κατανοητό τρόπο.

Στο πέμπτο κεφάλαιο, με το οποίο ολοκληρώνεται η παρούσα πτυχιακή εργασία, γίνεται μια εφαρμογή δεδομένων με χρήση του Προγράμματος Στατιστικής SPSS, που αποτελεί ένα από τα πιο ευρέως αναγνωρισμένα και αξιόπιστα προγράμματα για τη στατιστική ανάλυση δεδομένων.

ΚΕΦΑΛΑΙΟ 1

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

1.1 Εισαγωγη

Η παλινδρόμηση είναι από τα πιο σημαντικά εργαλεία του αναλυτή για να αναλύσει τα οικονομικά και χρηματοοικονομικά φαινόμενα. Ασχολείται με την περιγραφή και αξιολόγηση των σχέσεων μεταξύ μιας μεταβλητής, η οποία καλείται εξαρτημένη (dependent) ή μεταβλητή απόκρισης (response) ή προβλέψιμη (predicted), και μιας ή περισσότερων μεταβλητών οι οποίες ονομάζονται ανεξάρτητες (independent) ή προβλεπτικές (predictive) ή επεξηγηματικές (explanatory). Η ανεξάρτητη μεταβλητή παίρνει το όνομα της καθόσον ελέγχεται με μετρήσεις που διεξάγει ο ερευνητής, το αποτέλεσμα των οποίων αναμένεται να διαπιστωθεί επί της εξαρτημένης μεταβλητής, της οποίας οι τιμές εξαρτώνται άμεσα από τις τιμές της πρώτης. Τέτοια εξαρτημένη σχέση καλείται παλινδρόμηση και πιο συγκεκριμένα όταν εμπλέκονται δύο μόνο μεταβλητές έχουμε την απλή παλινδρόμηση.

Σε πολλές στατιστικές εφαρμογές συναντάμε το πρόβλημα της μελέτης της σχέσης δύο ή περισσοτέρων τυχαίων μεταβλητών. Παραδείγματα τέτοιας σχέσης έχουμε στη μελέτη του ύψους και του βάρους μιας ομάδας ανθρώπων, του εισοδήματος και κατανάλωσης εργαζομένων σε μια εταιρεία κλπ. Το ζήτημα είναι αφ’ ενός να αποφασίσουμε αν υφίσταται μια τέτοια σχέση και στην περίπτωση που όντως υφίσταται να προσδιορίσουμε τη σχέση. Ένας από τους κύριους λόγους που η μελέτη αυτή είναι σημαντική, κυρίως σε εφαρμογές που έχουν σχέση με επιχειρήσεις και με την οικονομία, είναι ότι οι σχέσεις αυτές χρησιμοποιούνται συχνά για προβλέψεις. Είναι προφανές ότι συχνά, είτε ιδιωτικές εταιρείες είτε κρατικές μονάδες χρειάζεται να προβλέψουν ποσότητες/χαρακτηριστικά όπως η ζήτηση, τα επιτόκια, ο πληθωρισμός, οι τιμές πρώτων υλών, το εργατικό κόστος κλπ. Είναι ενδιαφέρον λοιπόν να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν σε κάποιες άλλες μεταβλητές.

Η (γραμμική) παλινδρόμηση αποτελεί μια στατιστική μέθοδο η οποία αποσκοπεί στον προσδιορισμό ενός μαθηματικού μοντέλου για την περιγραφή, ερμηνεία, πρόβλεψη των τιμών ενός χαρακτηριστικού (μεταβλητής) σε σχέση με τις τιμές ενός πλήθους άλλων χαρακτηριστικών (μεταβλητών). Αρχικά θα ασχοληθούμε με την απλούστερη περίπτωση παλινδρόμησης που είναι η απλή γραμμική παλινδρόμηση, κατά την οποία υπάρχει μία μόνο ανεξάρτητη μεταβλητή Χ και η εξαρτημένη μεταβλητή Υ που μπορεί να προσεγγιστεί ικανοποιητικά από μια γραμμική συνάρτηση του Χ. Η σχέση μεταξύ των δύο μεταβλητών χαρακτηρίζεται ως αιτιώδης διότι οι τιμές των ερμηνευτικών μεταβλητών, ερμηνεύουν την τιμή της εξαρτημένης. Η απλή σχέση παλινδρόμησης μπορεί να παρουσιαστεί με την εξής μορφή: *Y=f(X)*.

Ανεξάρτητα, όμως, από τους λόγους για τους οποίους η μελέτη της σχέσης δύο ή περισσοτέρων μεταβλητών είναι χρήσιμη, το πρώτο βήμα για να πραγματοποιηθεί η μελέτη αυτή είναι η κατασκευή μιας μαθηματικής εξίσωσης (μοντέλου) που περιγράφει τη φύση της σχέσης που υφίσταται μεταξύ των υπό μελέτη μεταβλητών. Η διαδικασία δημιουργίας μιας μαθηματικής "εξίσωσης" για την περιγραφή ενός φαινομένου μπορεί να είναι ιδιαίτερα πολύπλοκη. Αυτό οφείλεται στο γεγονός ότι για την κατασκευή του μοντέλου απαιτείται κάποια γνώση της φύσης της σχέσης μεταξύ των μεταβλητών. Για παράδειγμα, ένας επενδυτής προκειμένου να προβεί στην αγορά μετοχών θα ενδιαφερόταν να προβλέψει την μελλοντική τιμή των μετοχών. Παράγοντες που μπορούν να επηρεάσουν την τιμή αυτή είναι τα καθαρά έσοδα μιας εταιρείας, η ζήτηση κτλ. Έτσι, αν ο επενδυτής υποθέσει ότι υπάρχει θετική γραμμική σχέση μεταξύ των μεταβλητών, αυτό θα συνεπάγεται ότι μία αύξηση των καθαρών εσόδων της εταιρείας, μπορεί να οδηγήσει και σε αντίστοιχη αύξηση της τιμής της μετοχής.

Η σχέση που συνδέει την εξαρτημένη μεταβλητή με τις ανεξάρτητες είναι στατιστική και όχι συναρτησιακή. Στην στατιστική σχέση, για κάθε τιμή της ανεξάρτητης μεταβλητής υπολογίζεται μια θεωρητική τιμή της εξαρτημένης μεταβλητής, ενώ η πραγματική τιμή της βρίσκεται μέσα σε ένα εύρος τιμών το οποίο περιέχει την θεωρητική τιμή. Στην συναρτησιακή σχέση, δηλαδή σε μια εξίσωση, κάθε τιμή της ανεξάρτητης μεταβλητής δίνει πάντα την ίδια τιμή στην εξαρτημένη μεταβλητή (μορφή Y=f(X)). Ωστόσο, για ευκολία χρησιμοποιούμε

τον όρο "εξισώσεις παλινδρόμησης", παρόλο που δεν πρόκειται για εξίσωση, αλλά για στατιστικό μοντέλο.

Έστω ότι έχουμε δύο μεταβλητές Χ και Υ. Στα μαθηματικά έχουμε συναρτησιακή σχέση μεταξύ των μεταβλητών της μορφής $Y = f(X)$ ή $Y = \beta_0 + \beta_1 X$ στο γραμμικό υπόδειγμα, επισημαίνοντας ότι ο όρος "γραμμικό" για τον χαρακτηρισμό του υποδείγματος αναφέρεται στις παραμέτρους και όχι στις μεταβλητές.

Σε διάφορες επιστήμες συναντώνται συχνά σχέσεις μεταξύ χαρακτηριστικών. Για παράδειγμα, σε μαθήματα οικονομικού περιεχομένου συναντώνται επίσης σχέσεις όπως οι εξής:

- Κέρδος = Εισπράξεις – Κόστος
- Συνολικό κόστος = Σταθερό κόστος + (Μεταβλητό κόστος · αριθμό μονάδων που παρήχθησαν).

Στην καθημερινή ζωή, ωστόσο, τα πράγματα δεν είναι πάντα ιδεώδη, είναι δηλαδή σχεδόν απίθανο να έχουμε δύο μεγέθη που να έχουν μια τέλεια μαθηματική σχέση. Τα σφάλματα μετρήσεως είναι αναπόφευκτα, επομένως ακόμα και αν η θεωρητική σχέση ανάμεσα στις μεταβλητές είναι ακριβής, πάλι θα υπάρχουν αποκλίσεις από τη θεωρητική σχέση, που θα οφείλονται στην ύπαρξη λαθών στην μέτρηση των τιμών των μεταβλητών. Στις περισσότερες επομένως περιπτώσεις που αναφέρονται σε πρακτικά προβλήματα, πρέπει να χρησιμοποιηθούν μοντέλα που να περικλείουν το στοιχείο της τυχαιότητας, στοιχείο που είναι μέρος της καθημερινής ζωής.

1.2 ΜΕΛΕΤΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΗΣ ΑΠΛΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Όπως προαναφέρθηκε, η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση, κατά την οποία υπάρχει μόνο μια **ανεξάρτητη μεταβλητή** *X*, και η **εξαρτημένη μεταβλητή** *Y*, η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του Χ. Η περίπτωση αυτή εμφανίζεται τόσο σε πειραματικές όσο και σε μη πειραματικές μελέτες. Στις πειραματικές μελέτες ο ερευνητής καθορίζει, για παράδειγμα, από πριν τις δόσεις ενός φαρμάκου (ανεξάρτητη μεταβλητή) που δίνει στα πειραματόζωα και μετρά τις αντιδράσεις τους (εξαρτημένη μεταβλητή). Με την παλινδρόμηση ενδιαφέρεται να προσδιορίσει μία σχέση δόσης-αντίδρασης για το συγκεκριμένο φάρμακο. Στις μη πειραματικές μελέτες ή δειγματοληψίες, γίνονται μετρήσεις σε δύο χαρακτηριστικά (μεταβλητές) για κάθε άτομο (μονάδα) του δείγματος. Σε ένα δείγμα 10 μαθητών μετράμε, για παράδειγμα, το βάρος και το ύψος τους. Η διάκριση εδώ μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής είναι δύσκολη. Αν αυτό που μας ενδιαφέρει είναι το "τι συμβαίνει με το βάρος των παιδιών όταν αλλάζει το ύψος τους", τότε θεωρούμε ως ανεξάρτητη μεταβλητή Χ το ύψος και ως εξαρτημένη μεταβλητή Υ το βάρος. Οπότε, ενδιαφερόμαστε για την **παλινδρόμηση του βάρους (Υ) πάνω στο ύψος (Χ)**. Αντίθετα, αν μας ενδιαφέρει το "τι συμβαίνει με το ύψος των παιδιών όταν αλλάζει το βάρος τους", τότε θεωρούμε ως ανεξάρτητη μεταβλητή Χ το βάρος και ως εξαρτημένη μεταβλητή Υ το ύψος. Τότε έχουμε παλινδρόμηση του ύψους (Υ) πάνω στο βάρος (Χ).

1.2.1 ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ

Ο παρακάτω πίνακας 1 δίνει τα ύψη *X* (σε cm) και τα βάρη *Y* (σε kg) των 18 μαθητών της Γ΄ Λυκείου. Οι τιμές του ύψους δίνονται σε αύξουσα σειρά.

Πίνακας 1[1]

Λίστα υψών (σε cm) και βαρών (σε kg) 18 μαθητών.

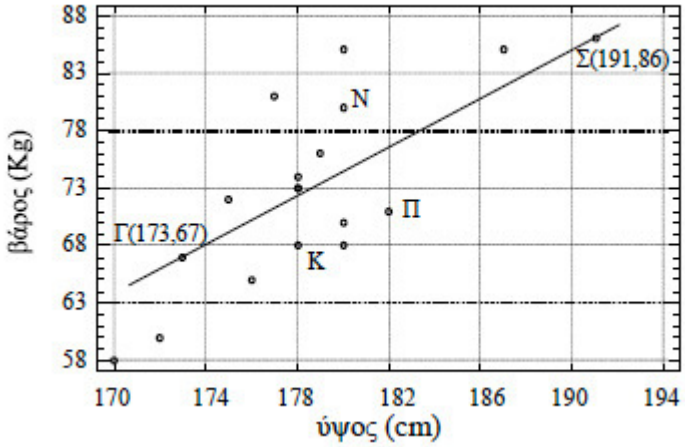
Μαθητής	Ύψος *X*	Βάρος *Y*	Μαθητής	Ύψος *X*	Βάρος *Y*
Α	170	58	Κ	178	68
Β	172	60	Λ	179	76
Γ	173	67	Μ	180	68
Δ	175	72	Ν	180	80
Ε	176	65	Ξ	180	70
Ζ	177	81	Ο	180	85
Η	178	73	Π	182	71
Θ	178	74	Ρ	187	85
Ι	178	73	Σ	191	86

Στο παράδειγμα αυτό έχουμε την περίπτωση όπου σε κάθε άτομο (μαθητή) γίνονται δύο μετρήσεις. Δηλαδή το δείγμα αποτελείται από τα ζεύγη τιμών των συνεχών μεταβλητών *X* (ύψος) και *Y* (βάρος).

Αν παραστήσουμε τα ζεύγη (*x*, *y*) των παρατηρήσεων σε ένα σύστημα ορθογώνιων αξόνων, παρατηρούμε ότι προκύπτει μία "διασπορά" των σημείων που αντιστοιχούν στους μαθητές που εξετάζουμε. Η παράσταση αυτή των σημείων καλείται **διάγραμμα διασποράς**, βλέπε σχήμα 1. Εδώ στον άξονα *χ* απεικονίζεται ως ανεξάρτητη μεταβλητή το ύψος *X* και στον άξονα y ως εξαρτημένη μεταβλητή το βάρος *Y*.

[1] Βλ. ΜΑΘΗΜΑΤΙΚΑ ΚΑΙ ΣΤΟΙΧΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ, Βιβλίο Μαθητή Γ΄ Γενικού Λυκείου Γενικής Παιδείας, ΙΝΣΤΙΤΟΥΤΟ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΕΚΔΟΣΕΩΝ «ΔΙΟΦΑΝΤΟΣ», κεφ. 2.4 Γραμμική Παλινδρόμηση, σελ. 106.

Σχήμα 1



Διάγραμμα διασποράς και ευθεία προσαρμοσμένη "με το μάτι" για τα δεδομένα του πίνακα 1.

Η προσεκτική παρατήρηση ενός διαγράμματος διασποράς μπορεί να μας δώσει σημαντικές πληροφορίες για τη σχέση εξάρτησης που ενδεχομένως υπάρχει μεταξύ των μεταβλητών τις οποίες εξετάζουμε. Η πείρα μας λέει ότι υπάρχει κάποια σχέση μεταξύ του ύψους και του βάρους κάθε μαθητή και μάλιστα θετική. Όταν αυξάνεται το ύψος, τότε αυξάνει και το βάρος. Στο παράδειγμα αυτό το διάγραμμα διασποράς δείχνει, γενικά, ότι οι ψηλοί μαθητές είναι συνήθως και πιο βαρείς. Για παράδειγμα, ο *Ν* είναι ψηλότερος και βαρύτερος από τον *Κ*, ο *Π* είναι ψηλότερος και βαρύτερος από τον *Κ*, αλλά βαρύτερος από τον *Π*.

1.2.2 Ευθεια Παλινδρομησης

Από το διάγραμμα διασποράς του προηγούμενου παραδείγματος φαίνεται καθαρά ότι υπάρχει μία σχέση ανάμεσα στο ύψος *X* και το βάρος *Y* των 18 μαθητών της Γ' Λυκείου. Τα σημεία (*x*, *y*) είναι συγκεντρωμένα περίπου γύρω από μία ευθεία, δηλαδή η σχέση μεταξύ των *X* και *Y* είναι κατά προσέγγιση γραμμική. Όπως έχουμε ήδη αναφέρει, μπορούμε να θεωρήσουμε τη μία μεταβλητή ως ανεξάρτητη μεταβλητή και την άλλη ως εξαρτημένη. Εδώ θεωρούμε ως ανεξάρτητη μεταβλητή το ύψος *X* και ως εξαρτημένη μεταβλητή το βάρος *Y*, οπότε η ευθεία που θα προσαρμόζεται καλύτερα στα σημεία αυτά καλείται **ευθεία παλινδρόμησης της *Y* πάνω στη *X*.**

Όπως γνωρίζουμε, η εξίσωση μιας ευθείας δίνεται από τη σχέση:

$$y = \alpha + \beta x \qquad (1)$$

όπου α και β είναι παράμετροι τις οποίες θέλουμε να υπολογίσουμε ή, όπως λέμε, να "εκτιμήσουμε", έτσι ώστε η ευθεία που θα προκύψει να μας δίνει όσο το δυνατόν την καλύτερη περιγραφή της σχέσης (εξάρτησης) που υπάρχει μεταξύ των μεταβλητών *X* και *Y*.

Η παράμετρος α μας δίνει το σημείο $(0,\alpha)$, όπου η ευθεία αυτή τέμνει τον άξονα *y'y*, ενώ η παράμετρος β παριστάνει το συντελεστή διεύθυνσης της ευθείας.

Ο πιο εύκολος τρόπος χάραξης της ευθείας είναι αυτός που γίνεται "με το μάτι". Μια τέτοια ευθεία έχουμε φέρει και στο διάγραμμα διασποράς του σχήματος 1. Για να βρούμε τα α και β, εργαζόμαστε ως εξής:

- Επιλέγουμε δύο σημεία, έστω τα *Γ*(173,67) και *Σ*(191,86) πάνω στην ευθεία που φέραμε "με το μάτι".
- Αντικαθιστούμε τις συντεταγμένες (x, y) των σημείων αυτών στην (1), οπότε προκύπτει το σύστημα:

$$\begin{cases} y_1 = \alpha + \beta x_1 \\ y_2 = \alpha + \beta x_2 \end{cases} \Leftrightarrow \begin{cases} 67 = \alpha + 173\beta \\ 86 = \alpha + 191\beta \end{cases}.$$

- Επιλύοντας το σύστημα αυτό βρίσκουμε α = -115,6 και β = 1,06 οπότε η εξίσωση της ευθείας (1) γίνεται:

$$y = -115{,}6 + 1{,}06 \qquad (2)$$

Επομένως, η ευθεία που κατά τη γνώμη μας προσαρμόζεται καλύτερα στα σημεία του διαγράμματος διασποράς διέρχεται από το σημείο (0 , -115,6) και έχει συντελεστή διεύθυνσης 1,06.

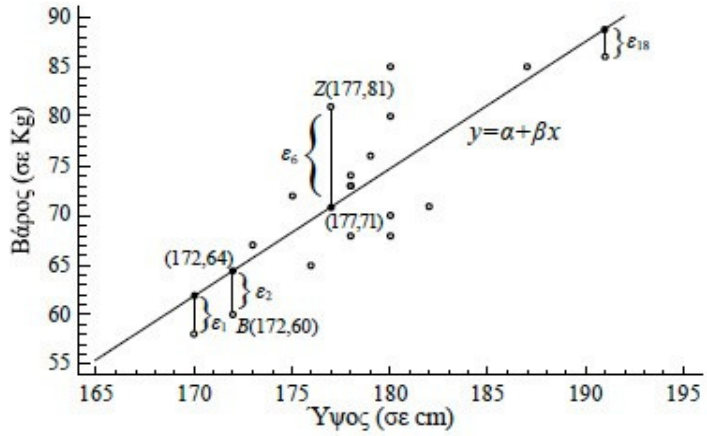
1.2.3 Εκτιμηση Της Ευθειας Παλινδρομησης Με Τη Μεθοδο Των Ελαχιστων Τετραγωνων

Είδαμε ότι η πιο απλή διαδικασία προσαρμογής μιας ευθείας γραμμής σε ένα διάγραμμα διασποράς είναι "με το μάτι". Αυτή όμως έχει πολλά μειονεκτήματα παρά την απλότητά της. Το κυριότερο είναι η έλλειψη αντικειμενικότητας, αφού διάφορα άτομα μπορούν να χαράξουν διαφορετικές μεταξύ τους ευθείες. Ακόμα και το ίδιο άτομο μπορεί να χαράζει διαφορετικές ευθείες κάθε φορά.

Χρειαζόμαστε λοιπόν μια ακριβέστερη μέθοδο για την προσαρμογή μιας ευθείας γραμμής σε τέτοιου είδους δεδομένα.
Μια μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων α και β, άρα και για την εύρεση της εξίσωσης της καλύτερης ευθείας που προσαρμόζεται στα δεδομένα, είναι η "**μέθοδος ελαχίστων τετραγώνων**". Η πρώτη αναφορά με ολοκληρωμένη ανάπτυξη της μεθόδου των ελαχίστων τετραγώνων εμφανίζεται το 1805 σε μια εργασία του Γάλλου μαθηματικού Legendre, (1752-1833) και αμέσως μετά από το Γερμανό μαθηματικό Gauss, (1777-1855) στην αστρονομική του πραγματεία "Theoria Motus" για τον προσδιορισμό της τροχιάς του μικρού πλανήτη Δήμητρα.

Μάλιστα εδώ ο Gauss αναφέρει ότι χρησιμοποίησε την αρχή των ελαχίστων τετραγώνων πριν από το 1794 (σε ηλικία μόλις 17 ετών), έτσι ώστε να προηγείται του Legendre ως προς την ανακάλυψη αυτής της μεθόδου. Ας δούμε ξανά το διάγραμμα διασποράς στο σχήμα 2 του προηγούμενου παραδείγματος για τα ύψη X και τα βάρη Y των 18 μαθητών του πίνακα 1. Στο διάγραμμα αυτό έχουμε φέρει και μία ευθεία $y = \alpha + \beta x$, που πιστεύουμε ότι προσαρμόζεται καλύτερα στα σημεία (x_i , y_i) για τις $v = 18$ συνολικά μετρήσεις των μεταβλητών X και Y.

Σχήμα 2



Προσαρμογή ευθείας ελαχίστων τετραγώνων στο διάγραμμα διασποράς των δεδομένων του πίνακα 1.

Έτσι, για παράδειγμα, για το μαθητή B, σημείο $B(172,60)$, με ύψος $x_2 = 172$ cm έχουμε βρει, όπως φαίνεται στον πίνακα 1, βάρος $y_2 = 60$kg, ενώ, σύμφωνα με την ευθεία που φέραμε, το βάρος του αναμένεται να είναι (περίπου) 64kg,

έχουμε δηλαδή ένα σφάλμα $\varepsilon_2 = 60 - 64 = -4$, δηλαδή βάρος 4kg λιγότερο από το αναμενόμενο. Ομοίως για το μαθητή *Z*, σημείο *Z*(177,81), το βάρος του που μετρήθηκε ήταν $y_6 = 81$kg, ενώ το αναμενόμενο βάρος του σύμφωνα με την ευθεία που φέραμε είναι 71kg, έχουμε δηλαδή ένα σφάλμα $\varepsilon_6 = 81 - 71 = 10$, δηλαδή βάρος 10kg περισσότερο από το αναμενόμενο. Ανάλογα σφάλματα υπολογίζονται και για τους άλλους μαθητές. Θα θέλαμε λοιπόν να βρούμε με κάποια μέθοδο εκείνη την ευθεία $y = \alpha + \beta x$, έτσι ώστε τα σφάλματα που προκύπτουν να είναι όσο το δυνατόν μικρότερα.

Η μέθοδος των ελαχίστων τετραγώνων συνίσταται στον προσδιορισμό των παραμέτρων α, β, έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (x_i , y_i) από την ευθεία $y = \alpha + \beta x$, δηλαδή το

$$\sum_{i=1}^{v} \varepsilon_i^2 = \sum_{i=1}^{v} (y_i - \alpha - \beta x_i)^2 \qquad (3)$$

να γίνεται ελάχιστο.

Οι τιμές των παραμέτρων α και β, που ελαχιστοποιούν την (3), καλούνται **εκτιμήτριες ελαχίστων τετραγώνων**, συμβολίζονται με $\hat{\alpha}$ ("α καπέλο") και $\hat{\beta}$ ("β καπέλο"), αντιστοίχως, και αποδεικνύεται (η απόδειξη εδώ παραλείπεται) ότι δίνονται από τις σχέσεις:

$$\hat{\beta} = \frac{v\sum_{i=1}^{v} x_i y_i - \left(\sum_{i=1}^{v} x_i\right)\left(\sum_{i=1}^{v} y_i\right)}{v\sum_{i=1}^{v} x_i^2 - \left(\sum_{i=1}^{v} x_i\right)^2} \qquad (4)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Όπου $\bar{y} = \frac{1}{v}\sum_{i=1}^{v} y_i , \quad \bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i .$

Η ευθεία

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \qquad (5)$$

Καλείται ευθεία **ελαχίστων τετραγώνων** ή **ευθεία παλινδρόμησης** της *Y* (πάνω) στη *X*. Αντικαθιστώντας το $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ στη σχέση (5) βρίσκουμε την

$$\hat{y} - \bar{y} = \hat{\beta}(x - \bar{x}) ,$$

η οποία φανερώνει ότι η ευθεία ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ διέρχεται από το σημείο με συντεταγμένες $(\bar{x}, \bar{y})$ και έχει συντελεστή διεύθυνσης το $\hat{\beta}$. Αντικαθιστώντας τις τιμές x_i και y_i από τον πίνακα 1 στις σχέσεις (4) βρίσκουμε:

$$\hat{\beta} = 1{,}28 \quad \text{και} \quad \hat{\alpha} = -156{,}1$$

οπότε η ευθεία ελαχίστων τετραγώνων που προσαρμόζεται καλύτερα στα δεδομένα είναι από τη σχέση (5), η

$$\hat{y} = -156{,}1 + 1{,}28x .$$

Παρατηρούμε ότι υπάρχει σημαντική διαφορά από την ευθεία $y = -115{,}6 + 1{,}06x$ που προσαρμόσαμε "με το μάτι" στο σχήμα 1.

1.2.4 Ερμηνεια και Ιδιοτητες Εκτιμητων Ελαχιστων Τετραγωνων

Στην εξίσωση ελαχίστων τετραγώνων $\hat{y} = \hat{\alpha} + \hat{\beta}x$ η τιμή της εκτιμήτριας $\hat{\alpha}$ της παραμέτρου *α* παριστάνει την τεταγμένη του σημείου στο οποίο η ευθεία τέμνει τον άξονα *y'y*, δηλαδή την τιμή της εξαρτημένης μεταβλητής *Y* όταν $x = 0$. Όταν το $\hat{\alpha} = 0$ τότε η ευθεία διέρχεται από την αρχή των αξόνων.

Έστω τώρα δυο τιμές x_1 και $x_2 = x_1 + 1$ της ανεξάρτητης μεταβλητής. Τότε λαμβάνοντας τη διαφορά των αντίστοιχων προβλεπόμενων τιμών της εξαρτημένης μεταβλητής βρίσκουμε

$$\hat{y}_2 - \hat{y}_1 = (\hat{\alpha} + \hat{\beta}x_2) - (\hat{\alpha} + \hat{\beta}x_1) = \hat{\alpha} + \hat{\beta}(x_1 + 1) - (\hat{\alpha} + \hat{\beta}x_1) = \hat{\beta}$$

δηλαδή $\widehat{y_2} = \widehat{y_1} + \hat{\beta}$. Συνεπώς ο συντελεστής διεύθυνσης $\hat{\beta}$ της ευθείας $\hat{y} = \hat{\alpha} + \hat{\beta}x$ παριστά τη μεταβολή της εξαρτημένης μεταβλητής *Y* όταν το *X* μεταβληθεί κατά μια μονάδα. Έτσι, όταν το *x* αυξηθεί κατά μια μονάδα τότε το $\hat{y}$ αυξάνεται κατά $\hat{\beta}$ μονάδες όταν $\hat{\beta} > 0$ ή ελαττώνεται κατά $\hat{\beta}$ μονάδες όταν $\hat{\beta} < 0$.

Σύμφωνα με το θεώρημα των Gauss-Markov, για το κλασικό γραμμικό υπόδειγμα, οι εκτιμητές που προκύπτουν από τη μέθοδο των ελαχίστων τετραγώνων είναι άριστοι, γραμμικοί και αμερόληπτοι εκτιμητές.

Θεώρημα των Gauss-Markov:
Για το απλό γραμμικό μοντέλο, οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\alpha}$ και $\hat{\beta}$ είναι:

1. γραμμικές συναρτήσεις των παρατηρήσεων της εξαρτημένης μεταβλητής Y,
2. αμερόληπτες,
3. μεταξύ όλων των γραμμικών αμερόληπτων εκτιμητών, έχουν την μικρότερη διακύμανση.

1.2.5 Συντελεστης Προσδιορισμου

Η διασπορά της μεταβλητής *Y* εκφράζεται με τις αποκλίσεις $y_i - \bar{y}$ των διαφόρων τιμών από τη μέση τιμή τους. Αν όλες οι τιμές ήταν ίσες μεταξύ τους δεν θα υπήρχε μεταβλητότητα στα δεδομένα και κάθε απόκλιση $y_i - \bar{y}$ θα ήταν ίση με το μηδέν. Όσο μεγαλύτερες είναι οι αποκλίσεις $(y_i - \bar{y})$, τόσο μεγαλύτερη θα είναι και η διασπορά των δεδομένων. Η ολική μεταβλητότητα (διασπορά) των παρατηρήσεων εκφράζεται σαν το άθροισμα των τετραγώνων των αποκλίσεων $(y_i - \bar{y})$ και συμβολίζεται με $\boldsymbol{SST = \sum(y_i - \bar{y})^2}$ που ονομάζεται **Ολικό άθροισμα τετραγώνων** (Total Sum of Squares).
Ένα νέο μέτρο της μεταβλητότητας των y_i γύρω από την ευθεία παλινδρόμησης, είναι το **άθροισμα τετραγώνων των σφαλμάτων** (Error Sum of Squares) $\sum {e_i}^2$ και συμβολίζεται με $\boldsymbol{SSE = \sum(y_i - \hat{y}_i)^2}$.
Επομένως η διαφορά των SST και SSE συμβολίζεται $\boldsymbol{SSR = SST - SSE}$.

Το $\boldsymbol{SSR = \sum(\hat{y}_i - \overline{y})^2}$ καλείται **άθροισμα τετραγώνων παλινδρόμησης** (Regression Sum of Squares) και εκφράζει την επίδραση της σχέσης παλινδρόμησης των δύο μεταβλητών στη μείωση της μεταβλητότητας των παρατηρήσεων y_i.

Γενικά αποδεικνύεται ότι ισχύει η σχέση $\boldsymbol{SST = SSR + SSE}$, δηλαδή η συνολική μεταβλητότητα των τιμών εκφράζεται σαν άθροισμα δύο όρων, της μεταβλητότητας που ερμηνεύεται από την παλινδρόμηση (SSR) και της μεταβλητότητας που παραμένει ανερμήνευτη από την μεταβλητή *X*, σαν το υπόλοιπο ή σφάλμα (SSE).

Το άθροισμα τετραγώνων που οφείλεται σε σφάλματα (SSE), μπορεί να ερμηνευθεί ως η ποσότητα μεταβλητότητας της *Y* που μένει ανερμήνευτη από το γραμμικό μοντέλο και ισχύει ότι $SSE \leq SST$, με την ισότητα να ισχύει όταν δεν υπάρχουν σφάλματα, δηλαδή όταν το μοντέλο μας περιγράφει με ακρίβεια την πληθυσμιακή εξίσωση παλινδρόμησης.

Η αναλογία της συνολικής διασποράς η οποία ερμηνεύεται από την παλινδρόμηση, ονομάζεται συντελεστής προσδιορισμού (coefficient of determination). Το μέτρο αυτό συμβολίζεται με R^2 και εκφράζει το ποσοστό της διασποράς της εξαρτημένης μεταβλητής *Y* που ερμηνεύεται από την ανεξάρτητη μεταβλητή *X*. Δηλαδή:

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n} \left(\widehat{\mathrm{Y}}_{\mathrm{i}-} \overline{\mathrm{Y}}\right)^2}{\sum_{i=1}^{n} (\mathrm{Y}_{\mathrm{i}-} \overline{\mathrm{Y}})^2} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

Η τιμή του συντελεστή προσδιορισμού R^2 κυμαίνεται από 0 έως 1. Αν το R^2 είναι κοντά στο 1, το μοντέλο της παλινδρόμησης έχει μεγάλη δυνατότητα ερμηνείας της εξαρτημένης μεταβλητής και τα σφάλματα είναι μικρά. Αντίθετα τιμές του R^2 κοντά στο 0, δείχνουν ότι δεν είναι επιτυχές το μοντέλο της παλινδρόμησης για την ερμηνεία της εξαρτημένης μεταβλητής. Ο συντελεστής προσδιορισμού παράγει αξιόπιστα αποτελέσματα όταν ο αριθμός των παρατηρήσεων είναι σημαντικά υψηλότερος από τον αριθμό των μεταβλητών. Μπορούμε όμως να υπολογίζουμε και την ποσότητα 1- R^2 που εκφράζει το ποσοστό της συνολικής μεταβλητότητας που οφείλεται σε σφάλματα, ανάλογα με το τι είναι αυτό που θέλουμε να μελετήσουμε. Ο συντελεστής προσδιορισμού αποτελεί σημειακή εκτίμηση (όχι αμερόληπτη) του πληθυσμιακού συντελεστή

προσδιορισμού και τείνει να τον υπερεκτιμά (θετικά ασύμμετρη η κατανομή του όταν το δείγμα είναι μικρό).

Ως μέτρο του βαθμού συσχετίσεως δύο τυχαίων μεταβλητών, χρησιμοποιείται ο συντελεστής συσχετίσεως ρ που ορίζεται ως το κλάσμα με αριθμητή την συνδιακύμανση των *X* και *Y*, και παρανομαστή το γινόμενο των τυπικών αποκλίσεων των *X* και *Y*. Ο συντελεστής ρ είναι μια άγνωστη παράμετρος του πληθυσμού, δηλαδή αναφέρεται σε μια παράμετρο της συνδυασμένης κατανομής των Χ και Υ. Ως εκτιμητής του ρ χρησιμοποιείται ο συντελεστής συσχετίσεως του δείγματος, που ορίζεται ως:

$$r = \frac{\sum(\mathrm{X} - \overline{\mathrm{X}})(\mathrm{Y} - \overline{\mathrm{Y}})}{\sqrt{\sum(\mathrm{X} - \overline{\mathrm{X}})^2}\sqrt{\sum(\mathrm{Y} - \overline{\mathrm{Y}})^2}}$$

Ισχύει ότι $r=\pm\sqrt{R^2}$, ωστόσο υπάρχει μεγάλη διαφορά στην ερμηνεία τους.

Ο συντελεστής συσχετίσεως του δείγματος *r* είναι ένας εκτιμητής του συντελεστή συσχετίσεως στον πληθυσμό ρ και δεν εξαρτάται από τις μονάδες μέτρησης των *X* και *Y* από την αρχή μέτρησης επάνω στους άξονες, είναι καθαρός αριθμός. Ισχύει ότι $-1 \leq r \leq +1$ με την τιμή -1 να σημαίνει ότι έχουμε πλήρη αρνητική γραμμική συσχέτιση και αντίστοιχα το +1 ότι έχουμε πλήρη θετική γραμμική συσχέτιση. Όταν $r = \pm 1$ η σχέση είναι αιτιοκρατική κι όχι πιθανοκρατική, γιατί γνωρίζοντας την τιμή της μιας τυχαίας μεταβλητής, γνωρίζουμε και την τιμή της άλλης τυχαίας μεταβλητής ακριβώς. Η μηδενική τιμή του συντελεστή συσχετίσεως μας δείχνει ότι δεν υπάρχει γραμμική συσχέτιση. Επίσης το ρ είναι μια άγνωστη παράμετρος της συνδυασμένης κατανομής δύο τυχαίων μεταβλητών, ενώ ο συντελεστής προσδιορισμού αναφέρεται στην αναλογία της μεταβλητότητας της *Y* που ερμηνεύει η μεταβλητή *X*, η οποία υποθέτουμε ότι δεν είναι τυχαία μεταβλητή. Επιπλέον, ο συντελεστής συσχετίσεως του δείγματος είναι μέτρο μόνο της γραμμικής συσχετίσεως ή εξαρτήσεως δύο μεταβλητών. Λόγω των ανωτέρω περιορισμών, καθώς και άλλων, η ανάλυση συσχετίσεως έχει περιορισμένη χρήση στην ανάλυση των οικονομικών δεδομένων.

1.2.6 Τα Σφαλματα Εκτιμησης η΄ Καταλοιπα

Έχοντας δημιουργήσει την γραφική παράσταση των μεταβλητών, η Μέθοδος Ελαχίστων Τετραγώνων χρησιμοποιείται για να επιλεχθεί η κατάλληλη γραμμή (από τις άπειρες) που περνάει από τα σημεία της γραφικής παράστασης των δύο αυτών μεταβλητών. Οι σχέσεις μεταξύ των μεταβλητών δεν είναι πάντα ακριβής. Μη παρατηρήσιμες ή τυχαίες διακυμάνσεις στα παρατηρηθέντα στοιχεία αναγκάζουν την αυστηρή μαθηματική σχέση μεταξύ των μεταβλητών να μην επαληθεύεται πάντα στην πράξη. Για να συμπεριληφθούν και οι συγκεκριμένες διακυμάνσεις, ένα στοχαστικό - τυχαίο τμήμα προστίθεται στο μοντέλο παλινδρόμησης. Αν γίνει χρήση της *X* για την επεξήγηση της συμπεριφοράς της *Y*, οποιαδήποτε ευθεία γραμμή μπορεί να αποδοθεί με τη μορφή:

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

Το $\beta_0+\beta_1 X$ είναι το συστηματικό τμήμα της εξίσωσης, ενώ το ε είναι το τυχαίο τμήμα, το οποίο ονομάζεται διαταρακτικός όρος (disturbance term) ή σφάλμα (error). Τα σφάλματα παρουσιάζονται στα πειράματα επειδή γίνονται λάθη κατά τη διαδικασία της μέτρησης της εξαρτημένης μεταβλητής ή επειδή το μοντέλο είναι ελλιπώς προσδιορισμένο. Η πρώτη αιτία είναι εύκολα κατανοητή. Η δεύτερη αιτία μπορεί να εξηγηθεί μέσω ενός παραδείγματος. Όπως είναι γνωστό, η κατανάλωση ενός αγαθού εξαρτάται όχι μόνο από την τιμή του συγκεκριμένου αγαθού αλλά και από πολλούς άλλους παράγοντες, όπως τις τιμές των υποκατάστατων και των συμπληρωματικών αγαθών, το εισόδημα, τα επιτόκια, το εισόδημα παρελθόντων χρόνων, το προσδοκώμενο μελλοντικό εισόδημα, την ηλικία του πληθυσμού, τη διαφήμιση, κλπ. Έτσι, αν προσδιοριστεί ένα υπόδειγμα κατανάλωσης ως γραμμική συνάρτηση της τιμής ή του εισοδήματος, αυτό δεν είναι επαρκές. Δύναται να υπάρχει η αντιμετώπιση ενός προβλήματος ελλιπούς προσδιορισμού. Όλοι οι παράγοντες, οι οποίοι επηρεάζουν την κατανάλωση και δεν συμπεριλήφθηκαν μέσα από το υπόδειγμα, θα αντιπροσωπεύονται από τον διαταρακτικό όρο ε_i.

Για να εκτιμηθεί ένα υπόδειγμα, πρέπει πρώτα να συλλεχθεί ένα δείγμα στοιχείων για την εξαρτημένη και την ανεξάρτητη μεταβλητή που ενδιαφέρει. Αν $Y_1,Y_2,\ldots,Y_n$ και $X_1,X_2,\ldots,X_n$ αντιπροσωπεύουν ένα τυχαίο δείγμα *n*

ανεξάρτητων παρατηρήσεων ενός πληθυσμού Y_i και X_i αντιπροσωπεύουν τις i^{th} τυχαίες παρατηρήσεις του δείγματος, τότε με δεδομένα τα n ζεύγη παρατηρήσεων Y_i και X_i, ο στόχος της ανάλυσης παλινδρόμησης είναι να αποκτηθούν εκτιμήσεις για τις άγνωστες πληθυσμιακές παραμέτρους β_0 και β_1. Πρακτικά όμως οι επιδράσεις στο τυχαίο τμήμα της παραπάνω εξίσωσης δεν μπορούν να προβλεφθούν. Είναι απαραίτητο να προσδιοριστεί μια κατανομή για τον διαταρακτικό όρο και να γίνει η υπόθεση για τα εξής:

1. Σε οποιαδήποτε τιμή της X, ο διαταρακτικός όρος είναι μία τυχαία μεταβλητή, η οποία κατανέμεται με μέσο 0 και διακύμανση σ^2. Δηλαδή $E(\varepsilon_i) = 0$ και $Var(\varepsilon_i) = \sigma^2$ για κάθε i. Δηλαδή ε_i είναι μια τυχαία μεταβλητή που παίρνει τιμές θετικές και αρνητικές έτσι ώστε η κατά μέσο όρο της να είναι μηδέν.
2. Οι δειγματικές τιμές του ε_i κατανέμονται ανεξάρτητα, δηλαδή τα σφάλματα δεν συσχετίζονται μεταξύ τους. Αυτό σημαίνει ότι για δύο διαφορετικές παρατηρήσεις του διαταρακτικού όρου ε_i και ε_j με $i \neq j$ η αναμενόμενη τιμή $E(\varepsilon_i, \varepsilon_j) = 0$ και η συνδιακύμανση τους (Cov) θα είναι μηδέν: $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i - E\varepsilon_j)(\varepsilon_j - E\varepsilon_i) = E\varepsilon_i, \varepsilon_j = 0$ καθώς $(E\varepsilon_i)=0$ και $(E\varepsilon_j)=0$. Όπως αναφέρθηκε στην 1η υπόθεση, κάθε δειγματικό σφάλμα ε_i κατανέμεται με την ίδια διακύμανση σ^2. Η διακύμανση της τυχαίας μεταβλητής είναι σταθερή για όλες τις τιμές της ανεξάρτητης μεταβλητής. Δηλαδή, η διασπορά των τιμών της ανεξάρτητης μεταβλητής. Σε αυτή την περίπτωση τονίζεται ότι ο όρος συμπεριφέρεται ομοσκεδαστικά.
3. Κάθε δειγματικό σφάλμα κατανέμεται κανονικά για κάθε i.
4. Οι τιμές της ανεξάρτητης μεταβλητής X λαμβάνονται ως σταθερές και για μία συγκεκριμένη τιμή της X αντιστοιχεί μια ολόκληρη κατανομή της Y. Έτσι κάθε διαφοροποίηση της Y οφείλεται στους παράγοντες που συμπεριλαμβάνονται στον διαταρακτικό όρο.

Οι υποθέσεις αυτές απαιτούν σωστό προσδιορισμό του υποδείγματος αναφορικά με τη συναρτησιακή μορφή και τις μεταβλητές που έχουν συμπεριληφθεί. Ο στόχος της ανάλυσης παλινδρόμησης είναι να εκτιμήσει τις παραμέτρους του υποδείγματος, ώστε η ανερμήνευτη μεταβολή της Y οριζόμενη ως το κατάλοιπο (residual), να είναι μικρή και μη συστηματική. Η παραβίαση των συγκεκριμένων υποθέσεων οδηγεί σε προβλήματα

αυτοσυσχέτισης, ετεροσκεδαστικότητας κλπ. Τα συγκεκριμένα προβλήματα αποτελούν σημαντικά θέματα και χρήζουν ιδιαίτερης ανάλυσης αν αποσκοπούν σε σωστή μοντελοποίηση κάποιων οικονομικών και κοινωνικών φαινομένων με σκοπό τις προβλέψεις.

Οι παραπάνω υποθέσεις υποδηλώνουν ότι τα σφάλματα είναι ανεξάρτητες μεταβλητές που κατανέμονται κανονικά ως $N(0,\sigma^2)$. Το άθροισμα των τετραγώνων των σφαλμάτων όπως προαναφέρθηκε έχει καθιερωθεί να συμβολίζεται με SSE. Μια ισοδύναμη έκφραση του SSE που διευκολύνει σημαντικά τους υπολογισμούς είναι η:

$$SSE = \sum \hat{\varepsilon}_i^{\,2} = \sum Y_i^{\,2} - \widehat{b_0} \sum Y_i - \widehat{b_1} \sum X_i Y_i$$

Βασικές ιδιότητες των σφαλμάτων είναι οι εξής:

1. Αποδεικνύεται ότι το SSE είναι το μικρότερο δυνατό (ελάχιστο).
2. Το άθροισμα των σφαλμάτων ισούται με μηδέν.
3. $\sum X_i \, \hat{\varepsilon}_i = 0$
4. $\sum \hat{Y}_i \, \hat{\varepsilon}_i = 0$

1.3 Εφαρμογη Της Αναλυσης Της Παλινδρομησης Στον Επενδυτικο Κινδυνο

Η Γραμμική Παλινδρόμηση χρησιμοποιείται για τη δημιουργία γραμμών τάσεως, χρησιμοποιώντας δεδομένα του παρελθόντος για να προβλέψει τις μελλοντικές αποδόσεις ή «τάσεις». Σε ένα επιχειρησιακό περιβάλλον η ανάλυση της παλινδρόμησης, χρησιμοποιείται συνήθως για να υποδειχθεί η κίνηση των οικονομικών ή τα χαρακτηριστικά ενός προϊόντος στη πάροδο του χρόνου. Σε ένα επενδυτικό πλαίσιο η ανάλυση της παλινδρόμησης χρησιμοποιείται ώστε να αναλυθούν οι τιμές των μετοχών, οι τιμές του πετρελαίου, ή οι προδιαγραφές ενός προϊόντος και έτσι να ορισθεί ένας σχετικός βαθμός επικινδυνότητας μιας επένδυσης. Έτσι η γραμμική παλινδρόμηση και η σωστή χρήση της θεωρείται το κλειδί για την εκτίμηση του κινδύνου που συνδέεται με τις περισσότερες επενδυτικές κινήσεις (Keener, 2011).

Για αυτό το λόγο με τη βοήθεια της ανάλυσης της γραμμικής παλινδρόμησης αναπτύχθηκε το μοντέλο Capital Asset Pricing το οποίο υπολογίζει ένα κοινό μέτρο της μεταβλητότητας μιας μετοχής ή επένδυσης το οποίο ονομάζεται beta (και καθορίζεται με τη βοήθεια της γραμμικής παλινδρόμησης). Στο χρηματοοικονομικό τομέα, το beta μιας μετοχής ή του χαρτοφυλακίου είναι ένας αριθμός που περιγράφει τη σχέση της απόδοσης της συγκεκριμένης μετοχής με αυτές της χρηματοπιστωτικής αγοράς στο σύνολό της (Levinson, 2006).

Ένα θετικό beta σημαίνει ότι οι αποδόσεις των περιουσιακών στοιχείων (ή μετοχών) ακολουθούν σε γενικές γραμμές τις αποδόσεις της σχετικής αγοράς, με την έννοια ότι είτε και οι δύο τείνουν να είναι ανώτερες των αντίστοιχων μέσων όρων τους μαζί, ή και τα δύο τείνουν να είναι κάτω των αντίστοιχων μέσων όρων τους από κοινού. Με την ίδια λογική ένα αρνητικό beta σημαίνει ότι οι αποδόσεις των περιουσιακών στοιχείων έχουν εν γένει αντίθετη κίνηση από τις αποδόσεις της αγοράς: όταν η τιμή της μίας είναι κάτω του μέσου όρου της, η άλλη θα τείνει να είναι άνω του μέσου όρου της (Myron & Joseph, 1977).

Σύμφωνα με τον McAlpine (2010) ο συντελεστής beta είναι μία βασική παράμετρος της στατιστικής διακύμανσης ενός περιουσιακού στοιχείου στο μοντέλο τιμολόγησης κεφαλαίου (CAPM) το οποίο δεν μπορεί να αφαιρεθεί από τη διαφοροποίηση που παρέχεται από τα χαρτοφυλάκια πολλών υψηλού κινδύνου περιουσιακών στοιχείων, λόγω της συσχέτισης των αποδόσεων του με τις αποδόσεις των άλλων περιουσιακών στοιχείων που βρίσκονται στο χαρτοφυλάκιο. Ο συντελεστής beta μπορεί να εκτιμηθεί για μεμονωμένες εταιρείες χρησιμοποιώντας την ανάλυση παλινδρόμησης έναντι ενός χρηματιστηριακού δείκτη. Ο τύπος για την beta ενός περιουσιακού στοιχείου σε ένα χαρτοφυλάκιο είναι:

$$\beta_a = \frac{\mathrm{Cov}(r_a, r_p)}{\mathrm{Var}(r_p)},$$

όπου το r_a μετρά το ποσοστό απόδοσης του περιουσιακού στοιχείου, r_p μετρά το ποσοστό απόδοσης του χαρτοφυλακίου, και Cov (r_a,r_p) είναι η συνδιακύμανση μεταξύ των ποσοστών της επιστροφής. Το χαρτοφυλάκιο του

ενδιαφέροντος στις διατυπώσεις του CAMP είναι το χαρτοφυλάκιο της αγοράς που περιλαμβάνει όλα τα επικίνδυνα στοιχεία ενεργητικών μετοχών, και έτσι οι όροι r_p του άνωθεν τύπου αντικαθίστανται από r_m (δηλαδή το ποσοστό απόδοσης της αγοράς).

Εξ ορισμού, ο συντελεστής beta της επενδυτικής αγοράς είναι 1.0 και οι μεμονωμένες μετοχές κατατάσσονται ανάλογα με το πόσο αποκλίνουν από τη μακροοικονομική της αγοράς (για λόγους απλούστευσης, η S&P 500 μερικές φορές χρησιμοποιείται ως υποκατάστατο για την αγορά στο σύνολό της). Μία μετοχή της οποίας η απόδοση είναι υψηλότερη από το μέσο όρο των αποδόσεων της αγοράς στην πάροδο του χρόνου μπορεί να έχει μια beta του οποίου η απόλυτη τιμή είναι μεγαλύτερη από 1.0 (αν είναι στην πραγματικότητα μεγαλύτερη από 1.0 θα εξαρτηθεί από το συσχετισμό των αποδόσεων της μετοχής με τις αποδόσεις της αγοράς). Ομοίως, μία μετοχή της οποίας η απόδοση είναι χαμηλότερη από το μέσο όρο των αποδόσεων της αγοράς έχει έναν beta με απόλυτη τιμή μικρότερη από 1.0 (Σαββίδης 1994). Μια μετοχή με beta 2 έχει επιστροφή (απόδοση) που αλλάζει κατά μέσο όρο δύο φορές συχνότερα από ότι οι αποδόσεις του συνόλου της αγοράς, έτσι για παράδειγμα όταν η απόδοση της αγοράς πέφτει ή ανεβαίνει κατά 3%, η απόδοση της μετοχής θα πέσει ή θα αυξηθεί (αντίστοιχα) κατά 6% σε μέσο όρο. Ωστόσο, επειδή ο συντελεστής beta εξαρτάται επίσης από τη συσχέτιση των αποδόσεων, μπορεί να υπάρχει σημαντική διακύμανση στο μέσο όρο: όσο υψηλότερος είναι ο συσχετισμός, τόσο μικρότερη η διακύμανση, και όσο χαμηλότερος είναι ο συσχετισμός, τόσο μεγαλύτερη είναι η διακύμανση). Επιπλέον, ο beta μπορεί επίσης να είναι και αρνητικός, που σημαίνει ότι οι αποδόσεις των μετοχών τείνουν να κινούνται προς την αντίθετη κατεύθυνση των αποδόσεων της αγοράς. Με αυτό τον τρόπο, μια μετοχή με beta -3 θα αντιμετωπίζει μείωση των αποδόσεων της 9% (κατά μέσο όρο) όταν η απόδοση της αγοράς αυξάνεται κατά 3%, και αντίστροφα, θα έχει άνοδο στις αποδόσεις της 9% (κατά μέσο όρο) εάν οι αποδόσεις της αγοράς μειωθούν κατά 3% (Tofallis, 2008).

Είναι ευρέως αποδεκτό πως, αν και οι μετοχές υψηλότερων beta παρέχουν τη δυνατότητα υψηλότερων αποδόσεων, τείνουν να είναι πιο ασταθείς και ως εκ

τούτου πιο ριψοκίνδυνες. Συνεπώς, οι μετοχές με χαμηλότερους συντελεστές beta είναι λιγότερο επιβλαβείς αλλά γενικά προσφέρουν χαμηλότερες αποδόσεις. Παρόλα αυτά, η ιδέα αυτή έχει αμφισβητηθεί από τον McAlpine (2010) ο οποίος υποστηρίζει ότι τα στοιχεία δείχνουν μικρή σχέση μεταξύ του beta και της πιθανής απόδοσης, και ότι τις περισσότερες φορές οι μετοχές χαμηλότερων beta είναι λιγότερο επικίνδυνες και περισσότερο κερδοφόρες. Τέλος, σύμφωνα με τον Klarman (1991), με τον ίδιο τρόπο που ο συντελεστής beta μιας μετοχής δείχνει τη σχέση της με αλλαγές της αγοράς, είναι επίσης και ένας δείκτης για την απαιτούμενη απόδοση των επενδύσεων (ROI). Λαμβάνοντας υπόψη ένα επιτόκιο μηδενικού κινδύνου ύψους 2%, για παράδειγμα, αν η αγορά (με beta 1) έχει αναμενόμενη απόδοση 8%, μία μετοχή με συντελεστή beta της τάξης του 1,5 θα πρέπει να έχει απόδοση 11% (= 2% + 1,5 (8% - 2%)).

ΚΕΦΑΛΑΙΟ 2

ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

2.1 Εισαγωγη

Σε πολλά πρακτικά προβλήματα είναι απαραίτητο να χρησιμοποιήσουμε δύο ή περισσότερες ανεξάρτητες μεταβλητές προκειμένου να ερμηνεύσουμε με μεγάλη ακρίβεια ένα φυσικό φαινόμενο και να βγάλουμε σωστότερα συμπεράσματα.

Για παράδειγμα, προκειμένου να χρησιμοποιηθεί ένα μοντέλο παλινδρόμησης για να προβλεφθεί η ζήτηση ενός προϊόντος μιας εταιρείας σε έναν αριθμό από διαφορετικές πόλεις, είναι ίσως σκόπιμο να χρησιμοποιηθούν κοινωνικοοικονομικές μεταβλητές (μέσο οικογενειακό εισόδημα, μόρφωση), δημογραφικές μεταβλητές (αριθμός μελών οικογένειας, αριθμός συνταξιούχων) και περιβαλλοντολογικές μεταβλητές (μέση ημερήσια θερμοκρασία) κ.α.

Όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές, για να ερμηνεύσουμε τη συμπεριφορά της εξαρτημένης μεταβλητής *Y*, χρησιμοποιούμε το μοντέλο της πολλαπλής παλινδρόμησης. Μάλιστα, αν η σχέση της εξαρτημένης μεταβλητής είναι γραμμική συνάρτηση των ανεξάρτητων μεταβλητών, τότε η περιγραφή της σχέσης αυτής γίνεται βάση ενός γραμμικού μοντέλου και έτσι αναφερόμαστε στην πολλαπλή γραμμική παλινδρόμηση.

Η πολλαπλή παλινδρόμηση έχει ευρεία επιστημονική αποδοχή, διότι θεωρείται ισχυρό και ευέλικτο στατιστικό εργαλείο με πλήθος εφαρμογών σε ποικίλα ερευνητικά πεδία. Κάποια από αυτά είναι:

- ✓ Διοίκηση επιχειρήσεων και έρευνα αγοράς: εκτίμηση του βαθμού επίδοσης του προσωπικού μιας εταιρείας, διαχείριση του αριθμού έκτασης των παραπόνων των πελατών.
- ✓ Προβλήματα οδικής συγκοινωνίας: διαχείριση του τύπου οδοστρώματος και είδους μεταφορικού μέσου στο χρόνο εκπλήρωσης μιας μετακίνησης.

- ✓ Υπέρβαση στον αθλητισμό: τρόποι βελτίωσης των αθλητικών επιδόσεων στο στίβο, προσαρμογή ενός βέλτιστου διαιτολογίου.
- ✓ Ατμοσφαιρική και υδρόβια ρύπανση με προεκτάσεις στη διαφύλαξη της δημόσιας υγείας.
- ✓ Τρόποι διερεύνησης της συμπεριφοράς του δείκτη νοημοσύνης σε διαγωνιστικό επίπεδο.
- ✓ Εκτίμηση της δράσης των χημικών συστατικών ενός τροφίμου στις οργανοληπτικές ιδιότητές του.

Συνοψίζοντας, η ανάλυση παλινδρόμησης χρησιμοποιείται για την περιγραφή των ειδικών σχέσεων μεταξύ των μεταβλητών, τη διακρίβωση θεωρητικών υποθέσεων, την πρόβλεψη από λήψεις πειραματικών δεδομένων και τη δημιουργία και επαλήθευση εξισώσεων πολλαπλής παλινδρόμησης.

2.2 ΜΕΛΕΤΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΗΣ ΠΟΛΛΑΠΛΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Όπως προαναφέρθηκε, μοντέλα παλινδρόμησης που περιέχουν δύο ή περισσότερες ανεξάρτητες μεταβλητές ονομάζονται μοντέλα πολλαπλής παλινδρόμησης (multiple regression models).

Αρχικά, θα μελετήσουμε το μοντέλο με δυο ανεξάρτητες μεταβλητές, το οποίο αποτελεί την φυσική επέκταση της απλής γραμμικής παλινδρόμησης ώστε να μελετώνται δυο ανεξάρτητες μεταβλητές X_1 και X_2. Έτσι θα έχουμε:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad i = 1,2, \ldots, n \tag{2.1}$$

όπου:

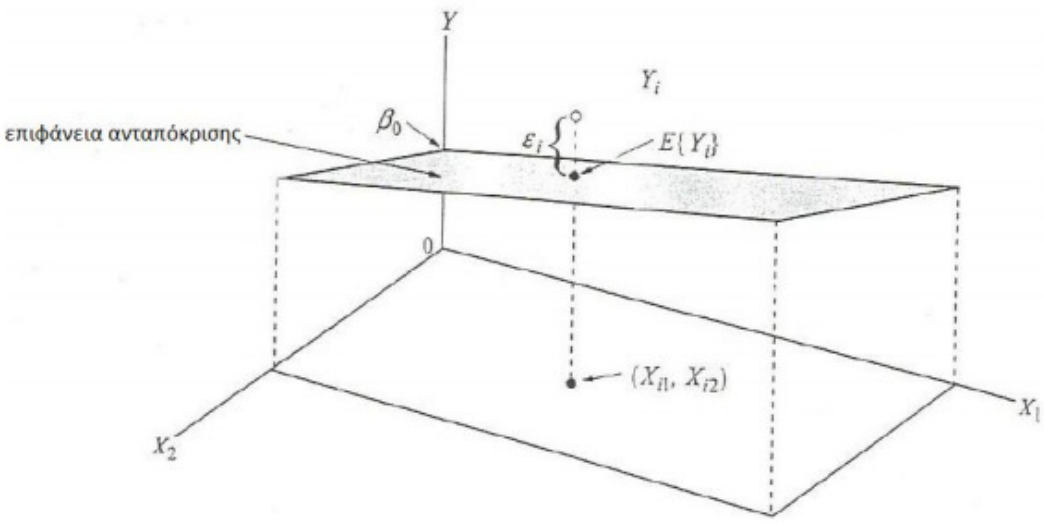
- Y_i είναι η τιμή της εξαρτημένης μεταβλητής στην i παρατήρηση.
- x_{i1}, x_{i2} είναι οι τιμές των ανεξαρτήτων μεταβλητών X_1και X_2 στην i παρατήρηση, οι οποίες υποτίθεται ότι είναι γνωστές.
- β_0, β_1 και β_2 είναι οι παράμετροι του μοντέλου.
- ε_i είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή $N(0, \sigma^2)$.

Επομένως, η συνάρτηση παλινδρόμησης (regression function) ή αλλιώς συνάρτηση ανταπόκρισης (response function) του μοντέλου (2.1) είναι $E(Y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Σε αυτό το σημείο αξίζει να σημειωθεί ότι η συνάρτηση αυτή ονομάζεται αρκετές φορές και επιφάνεια παλινδρόμησης (regression surface) ή επιφάνεια ανταπόκρισης (response surface).

Όπως και στην απλή γραμμική παλινδρόμηση, έτσι και στην πολλαπλή, οι παράμετροι έχουν ανάλογες ερμηνείες. Έτσι, στην επιφάνεια παλινδρόμησης:

- Το β_0 αντιστοιχεί στο σημείο τομής του άξονα του Y από την επιφάνεια (επίπεδο) παλινδρόμησης.
- Το β_1 δείχνει την μεταβολή της $E(Y)$ όταν το x_1 μεταβάλλεται κατά μια μονάδα ενώ το x_2 παραμένει σταθερό.
- Αντίστοιχα, το β_2 δείχνει την μεταβολή της $E(Y)$ όταν το x_2 μεταβάλλεται κατά μία μονάδα ενώ το x_1 παραμένει σταθερό.

Μια απεικόνισή της με την χρήση δύο ανεξάρτητων μεταβλητών, θα μπορούσε να είναι η παρακάτω:



Εικόνα 2.1

Αντίστοιχο του μοντέλου με δύο ανεξάρτητες μεταβλητές, είναι το μοντέλο παλινδρόμησης με p ανεξάρτητες μεταβλητές, το οποίο θα έχει τη μορφή:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad i = 1,\ldots, n \qquad (2.2)$$

όπου:

- Y_i είναι η τιμή της εξαρτημένης μεταβλητής για την i παρατήρηση.
- x_{i1} , x_{i2} , . . . , $x_{i,p-1}$ είναι οι τιμές των ανεξάρτητων μεταβλητών στην i παρατήρηση (υποτίθενται γνωστές σταθερές).
- β_i αντιπροσωπεύει την μεταβολή στην Y που προέρχεται από μια μεταβολή στην X_i κατά μία μονάδα, όταν όλες οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές.
- ε_i είναι ανεξάρτητες μεταβλητές που ακολουθούν την κατανομή $N(0, \sigma^2)$.

  Η ύπαρξη των καταλοίπων ε_i , όπως και στην απλή γραμμική παλινδρόμηση, είναι απαραίτητη γιατί στην πράξη κανένα μοντέλο δεν μπορεί να περιγράψει το σύνολο των πληροφοριών ενός σετ δεδομένων. Όσο καλά προσαρμοσμένη και να είναι η γραμμή πολλαπλής παλινδρόμησης στα δεδομένα, πάντα θα υπάρχει ένα μέρος της πληροφορίας που θα εξακολουθεί να μην ερμηνεύεται μέσω του

μοντέλου. Αυτός ο παράγοντας που δεν ερμηνεύεται από το γραμμικό μοντέλο καλείται λάθος της παλινδρόμησης.

Οπότε, η συνάρτηση παλινδρόμησης ή συνάρτηση ανταπόκρισης (η οποία μερικές φορές ονομάζεται και επιφάνεια παλινδρόμησης ή επιφάνεια ανταπόκρισης) είναι η:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1}, \quad \text{για i=1,....n.}$$

2.2.1 ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

Όπως και στο απλό γραμμικό μοντέλο, η δειγματική διασπορά των παρατηρήσεων Y_i χωρίζεται σε δύο αθροίσματα:

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2,$$

τα οποία συμβολίζονται και πάλι με SST, SSR και SSE αντίστοιχα. Το SST εκφράζει τη συνολική παρατηρούμενη μεταβλητότητα των Y_i, το SSR εκφράζει τη μεταβλητότητα των προσαρμοσμένων τιμών, ενώ το SSE εκφράζει τη μεταβλητότητα των Y_i σε σχέση με τις αντίστοιχες προσαρμοσμένες τιμές.

Ο πίνακας ανάλυσης διασποράς ANOVA είναι:

Πηγή Μεταβλητότητας	Αθροίσματα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Άθροισμα Τετραγώνων	Έλεγχος F
Παλινδρόμηση	$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	p-1	$MSR = \frac{SSR}{p-1}$	$F = \frac{MSR}{MSE}$
Υπόλοιπα	$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	n-p	$MSE = \frac{SSE}{n-p}$	
Σύνολο	$SST = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$	n-1		

Σημαντική διαφορά από τον πίνακα ανάλυσης διασποράς στην απλή παλινδρόμηση, είναι οι βαθμοί ελευθερίας. Η ποσότητα SST εξακολουθεί να

έχει n-1 βαθμούς ελευθερίας, ενώ η ποσότητα SSE έχει πλέον n-p βαθμούς ελευθερίας λόγω του ότι γίνεται εκτίμηση p μερικών συντελεστών παλινδρόμησης. Επίσης, η ποσότητα SSR έχει p-1 βαθμούς ελευθερίας που αντιπροσωπεύουν το πλήθος των μεταβλητών $X_1X_2...X_{p-1}$.

Στην πολλαπλή παλινδρόμηση το άθροισμα των τετραγώνων των υπολοίπων μπορεί να εκφραστεί ως:

$$SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}e_i^2=\underset{\sim}{e}'\underset{\sim}{e}$$

$$=\left(\underset{\sim}{Y}-X\underset{\sim}{b}\right)'\left(\underset{\sim}{Y}-X\underset{\sim}{b}\right)=\left(\underset{\sim}{Y}-H\underset{\sim}{Y}\right)'\left(\underset{\sim}{Y}-H\underset{\sim}{Y}\right)$$

$$=\underset{\sim}{Y}'\underset{\sim}{Y}-\underset{\sim}{Y}'H\underset{\sim}{Y}-\underset{\sim}{Y}'H'\underset{\sim}{Y}+\underset{\sim}{Y}'H'H\underset{\sim}{Y}$$

$$=\underset{\sim}{Y}'\underset{\sim}{Y}-\underset{\sim}{Y}'H\underset{\sim}{Y}=\underset{\sim}{Y}'(I_n-H)\underset{\sim}{Y}.$$

Ανάλογα:

$$SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\hat{Y}_i^2-2\bar{Y}\sum_{i=1}^{n}\hat{Y}_i+n\bar{Y}^2$$

$$=\sum_{i=1}^{n}\hat{Y}_i^2-n\bar{Y}^2=\underset{\sim}{\hat{Y}}'\underset{\sim}{\hat{Y}}-\frac{1}{n}\underset{\sim}{Y}'J_n\underset{\sim}{Y}=\left(H\underset{\sim}{Y}\right)'\left(H\underset{\sim}{Y}\right)-\frac{1}{n}\underset{\sim}{Y}'J_n\underset{\sim}{Y}$$

$$=\underset{\sim}{Y}'H\underset{\sim}{Y}-\frac{1}{n}\underset{\sim}{Y}'J_n\underset{\sim}{Y}=\underset{\sim}{Y}'\left(H-\frac{1}{n}J_n\right)\underset{\sim}{Y}$$

όπου **Jn=1*1^T** και **1** ο nx1 πίνακας ή πίνακας στήλη με στοιχεία του μονάδες. Επειδή SST=SSR+SSE, έχουμε πως:

$$SST=\underset{\sim}{Y}'\left(H-\frac{1}{n}J_n\right)\underset{\sim}{Y}+\underset{\sim}{Y}'(I_n-H)\underset{\sim}{Y}=\underset{\sim}{Y}'\left(I_n-\frac{1}{n}J_n\right)\underset{\sim}{Y}.$$

Αυτές οι εκφράσεις αποτελούν τις τετραγωνικές μορφές.

Όσον αφορά την ποσότητα MSE, όπως και στην απλή παλινδρόμηση, αποτελεί αμερόληπτο εκτιμητή της διακύμανσης σ^2.

Στην πολλαπλή παλινδρόμηση υπάρχουν και κάποια επιπλέον αθροίσματα τετραγώνων, τα οποία μετρούν την περιθώρια αύξηση στα αθροίσματα τετραγώνων της παλινδρόμησης όταν μία ή περισσότερες μεταβλητές προστίθενται στο μοντέλο παλινδρόμησης. Η περιθώρια αύξηση προσθέτοντας την X_2 σε ένα μοντέλο που ήδη έχει την X_1 , θα συμβολίζεται ως:

$$SSR(X_2 \setminus X_1) = SSR(X_1, X_2) - SSR(X_1),$$

το οποίο είναι ισοδύναμο με την έκφραση:

$$SSR(X_2 \setminus X_1) = SSE(X_1) - SSE(X_1, X_2).$$

Όταν στο μοντέλο υπάρχουν k μεταβλητές X, τότε θα υπάρχουν k αποσυνθέσεις των μεταβλητών X.

2.2.2 Μεθοδος Εκτιμησης Παραμετρων: Μεθοδος Ελαχιστων Τετραγωνων

Στην πολλαπλή παλινδρόμηση προσδιορίζονται περισσότερες παράμετροι με τρόπο ανάλογο όπως και για την απλή παλινδρόμηση. Όπως και εκεί, η γραμμή παλινδρομήσεως στον πληθυσμό είναι άγνωστη, εφόσον είναι άγνωστες οι τιμές των παραμέτρων β_i για i=0, 1,..., p-1. Για τον λόγο αυτό, θα πρέπει να εκτιμήσουμε τις τιμές των συντελεστών από ένα δείγμα παρατηρήσεων για τις μεταβλητές Y_i και X_{ij} [όπου η X_{ij} είναι η i (για i=1,...,n) παρατήρηση της j (για j=1,...,p-1) ανεξάρτητης μεταβλητής]. Με αυτό τον τρόπο κάνουμε μια εκτίμηση του πληθυσμιακού επιπέδου παλινδρόμησης από το δειγματικό

$E\left(\hat{Y}_i\right) = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_{p-1} X_{i,p-1}$ για i=1,...,n, με τα κατάλοιπα να ορίζονται ως το διάνυσμα $\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = \underset{\sim}{Y} - X\underset{\sim}{b}$.

Η **μέθοδος ελαχίστων τετραγώνων** για την εκτίμηση των παραμέτρων β_i για i=1,...,n, βασίζεται όπως και στο απλό γραμμικό μοντέλο στην ελαχιστοποίηση της παράστασης $\sum_{i=1}^{n} \varepsilon_i^2$. Έχουμε λοιπόν ότι:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \underset{\sim}{\varepsilon}' \underset{\sim}{\varepsilon} = \left(\underset{\sim}{Y} - X\underset{\sim}{\beta}\right)' \left(\underset{\sim}{Y} - X\underset{\sim}{\beta}\right)$$

$$= \left(\underset{\sim}{Y}' - \underset{\sim}{\beta}' X'\right)\left(\underset{\sim}{Y} - X\underset{\sim}{\beta}\right)$$

$$= \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{Y}' X\underset{\sim}{\beta} - \underset{\sim}{\beta}' X'\underset{\sim}{Y} + \underset{\sim}{\beta}' X' X\underset{\sim}{\beta}$$

$$= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}' X'\underset{\sim}{Y} + \underset{\sim}{\beta}' X' X\underset{\sim}{\beta}.$$

Ορίζουμε τις μερικές παραγώγους της σχέσης αυτής ως προς το διάνυσμα $\underline{\beta}$, οπότε:

$$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \underset{\sim}{\beta}} = \frac{\partial\left(\underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}' X'\underset{\sim}{Y} + \underset{\sim}{\beta}' X' X\underset{\sim}{\beta}\right)}{\partial \underset{\sim}{\beta}} = -2X'\underset{\sim}{Y} + 2X' X\underset{\sim}{\beta}$$

Θέτοντας τις μερικές παραγώγους ίσες με το μηδέν, οι κανονικές εξισώσεις ισορροπίας για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης είναι οι:

$$X' X\underset{\sim}{b} = X'\underset{\sim}{Y}$$

Σημειώνουμε ότι οι παράμετροι του διανύσματος $\underline{\beta}$ αντικαθίστανται από τις παραμέτρους του διανύσματος $\underline{b}$, καθώς οι δεύτερες αποτελούν εκτιμήτριες

των άλλοτε κανονικών εξισώσεων που ορίσαμε θέτοντας τις μερικές παραγώγους ίσες με μηδέν. Δηλαδή προκύπτει ότι οι εκτιμητές των συντελεστών $\underline{\beta}$ για το πολλαπλό γραμμικό μοντέλο παλινδρόμησης είναι οι:

$$\underset{\sim}{b} = \left(X^{'}X\right)^{-1} X^{'} \underset{\sim}{Y}.$$

Οι κανονικές εξισώσεις X′X $\underline{b}$ = X′$\underline{Y}$ στην αλγεβρική τους μορφή είναι:

$$\begin{pmatrix} 1 & 1\cdots & 1 \\ X_{11} & X_{21}\cdots & X_{n1} \\ \vdots & \vdots & \vdots \\ X_{1,p-1} & X_{2,p-1}\cdots & X_{n,p-1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} & X_{1,p-1} \\ 1 & X_{21} & X_{2,p-1} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n,p-1} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix} = \begin{pmatrix} 1 & 1\cdots & 1 \\ X_{11} & X_{21}\cdots & X_{n1} \\ \vdots & \vdots & \vdots \\ X_{1,p-1} & X_{2,p-1}\cdots & X_{n,p-1} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} n & \sum_{i=1}^{n} X_{i1}\cdots & \sum_{i=1}^{n} X_{i,p-1} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2\cdots & \sum_{i=1}^{n} X_{i1}X_{i,p-1} \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} X_{i,p-1} & \sum_{i=1}^{n} X_{p-1}X_{i1} & \sum_{i=1}^{n} X_{i,p-1}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{i,p-1}Y_i \end{pmatrix}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} Y_i = nb_0 + b_1\sum_{i=1}^{n} X_{i1} + \ldots + b_{p-1}\sum_{i=1}^{n} X_{i,p-1} \\ \sum_{i=1}^{n} X_{i1}Y_i = b_0\sum_{i=1}^{n} X_{i1} + b_1\sum_{i=1}^{n} X_{i1}^2 + \ldots + b_{p-1}\sum_{i=1}^{n} X_{i1}X_{i,p-1} \\ \vdots \\ \sum_{i=1}^{n} X_{i,p-1}Y_i = b_0\sum_{i=1}^{n} X_{i,p-1} + b_1\sum_{i=1}^{n} X_{p-1}X_{i1} + \ldots + b_{p-1}\sum_{i=1}^{n} X_{i,p-1}^2 \end{cases}$$

Το διάνυσμα των εκτιμητών των παραμέτρων για να υπολογιστεί, χρειάζεται τον πίνακα $(X\,X)^{-1}$. Για να αντιστρέφεται όμως ένας πίνακας, πρέπει να δειχθεί ότι η ορίζουσά του είναι διαφορετική του μηδενός.

2.2.3 Ιδιοτητες Των Εκτιμητων

Το **θεώρημα των Gauss-Markov**, το οποίο αναφέρεται στην αποτελεσματικότητα του εκτιμητή ελαχίστων τετραγώνων του γραμμικού μοντέλου παλινδρόμησης, διατυπώνει το εξής: Δεδομένων συγκεκριμένων υποθέσεων, ο εκτιμητής ελαχίστων τετραγώνων είναι αμερόληπτος και ο πιο αποτελεσματικός γραμμικός εκτιμητής των συντελεστών του μοντέλου γραμμικής παλινδρόμησης.

Επομένως, σύμφωνα με το θεώρημα των Gauss-Markov στο κλασικό γραμμικό υπόδειγμα οι εκτιμητές των συντελεστών $\underline{b}$ είναι γραμμικοί, αμερόληπτοι και άριστοι.

2.2.4 Συντελεστης Προσδιορισμου

Η αναλογία της συνολικής διασποράς η οποία ερμηνεύεται από την παλινδρόμηση, ονομάζεται συντελεστής προσδιορισμού και ισούται με το τετράγωνο του συντελεστή συσχέτισης του Pearson r. Το μέτρο αυτό συμβολίζεται με R^2 και αποτελεί ένα μέτρο του βαθμού προσαρμογής του επιπέδου παλινδρόμησης στις παρατηρήσεις του δείγματος. Ουσιαστικά μετράει την ερμηνευτική ικανότητα της εξίσωσης παλινδρόμησης και υπολογίζεται από τον τύπο:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Το πρόβλημα με τον συντελεστή πολλαπλού προσδιορισμού είναι ότι η τιμή του αυξάνει πάντα όταν αυξάνει ο αριθμός των ανεξάρτητων μεταβλητών, αφού προσθέτοντας ανεξάρτητες μεταβλητές βελτιώνουμε το μοντέλο με αποτέλεσμα να μειώνεται το άθροισμα των τετραγώνων των σφαλμάτων SSE, ενώ η ποσότητα SST παραμένει σταθερή. Προσθέτοντας όμως μια ανεξάρτητη μεταβλητή μπορεί να αυξάνουμε την τιμή του συντελεστή πολλαπλού προσδιορισμού, χάνουμε όμως έναν βαθμό ελευθερίας. Για τον λόγο αυτό χρησιμοποιούμε ένα τροποποιημένο μέτρο, τον διορθωμένο συντελεστή

πολλαπλού προσδιορισμού R_a^2 , που εκτός από την ποσότητα SSE λαμβάνει υπόψη και τους βαθμούς ελευθερίας. Ο υπολογισμός του διορθωμένου συντελεστή πολλαπλού προσδιορισμού γίνεται μέσω του τύπου:

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}.$$

Εκτός από τα μέτρα αυτά, υπάρχουν και οι μερικοί συντελεστές προσδιορισμού. Οι μερικοί συντελεστές προσδιορισμού μετρούν το καθαρό ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από την ανεξάρτητη μεταβλητή, απαλλαγμένη από την επίδραση άλλων μεταβλητών στο υπόδειγμα παλινδρόμησης.

Έτσι, στο πολλαπλό μοντέλο παλινδρόμησης με i μερικούς συντελεστές παλινδρομήσεως, ο μερικός συντελεστής προσδιορισμού ανάμεσα στην *Y* και την X_1 , υπό τον περιορισμό ότι οι υπόλοιπες X_i παραμένουν σταθερές, είναι:

$$r_{Y,1}^2 = \frac{\beta_1^2 \sum_{i=1}^{n}\left(X_{i1} - \bar{X}_1\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}.$$

Σε ένα μοντέλο με δύο ανεξάρτητες μεταβλητές *X*, μπορούμε να ορίσουμε τον μερικό συντελεστή προσδιορισμού μεταξύ της *Y* και της X_1 ενώ η X_2 είναι ήδη στο μοντέλο, να είναι:

$$r_{Y1.2}^2 = \frac{SSR\left(X_1 \setminus X_2\right)}{SSE\left(X_2\right)}$$

Ανάλογα, ο μερικός συντελεστής προσδιορισμού μεταξύ της *Y* και της X_2 ενώ η X_1 είναι ήδη στο μοντέλο, είναι:

$$r^2_{Y2.1} = \frac{SSR(X_2 \setminus X_1)}{SSE(X_1)}$$

Επιπλέον, σε ένα μοντέλο με τρεις ή περισσότερες μεταβλητές *X* μπορούμε να ορίσουμε μερικούς από τους παρακάτω μερικούς συντελεστές προσδιορισμού:

$$r^2_{Y1.23} = \frac{SSR(X_1 \setminus X_2, X_3)}{SSE(X_2, X_3)}$$

$$r^2_{Y2.13} = \frac{SSR(X_2 \setminus X_1, X_3)}{SSE(X_1, X_3)}$$

$$r^2_{Y4.123} = \frac{SSR(X_4 \setminus X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$ κ.ο.κ.

2.2.5 Επιλογη Μεταβλητων

Σε κάποια προβλήματα παλινδρόμησης, έχουμε στην διάθεσή μας δεδομένα από πολλούς παράγοντες που μπορεί να επηρεάζουν την εξαρτημένη μεταβλητή που μας ενδιαφέρει να καθορίσουμε ή να προβλέψουμε. Θα θέλαμε λοιπόν να επιλέξουμε το μικρότερο δυνατό υποσύνολο ανεξάρτητων μεταβλητών που εξηγεί το ίδιο καλά την εξαρτημένη μεταβλητή, όπως συμβαίνει και με τα μεγαλύτερα υποσύνολα ανεξάρτητων μεταβλητών ή ακόμα και ολόκληρο το σύνολο των ανεξάρτητων μεταβλητών.

Σε μια πρώτη προσέγγιση, η λύση είναι να βρούμε όλα τα δυνατά μοντέλα για όλους τους συνδυασμούς των ανεξάρτητων μεταβλητών και με βάση τον προσαρμοσμένο συντελεστή πολλαπλού προσδιορισμού, να βρούμε αυτό το μοντέλο που προσαρμόζεται καλύτερα. Αυτή η μέθοδος, αν και απλή στην σκέψη, δεν χρησιμοποιείται στην πράξη λόγω του μεγάλου αριθμού ανεξάρτητων μεταβλητών που μπορεί να έχουμε.

Εναλλακτικά, υπάρχουν κάποιες άλλες μέθοδοι που υπολογίζουν το βέλτιστο μοντέλο πολλαπλής παλινδρόμησης βηματικά και αυτές είναι:

1. Η μέθοδος απαλοιφής προς τα πίσω: Στην μέθοδο αυτή ξεκινάμε περιλαμβάνοντας όλες τις μεταβλητές στο μοντέλο και σε κάθε βήμα αποκλείεται μια μεταβλητή που δεν έχει σημαντική συνεισφορά σε αυτό. Η πρώτη μεταβλητή που αφαιρείται, είναι αυτή με το μικρότερο συντελεστή πολλαπλού προσδιορισμού R^2 και η διαδικασία συνεχίζεται έως ότου η αφαίρεση μεταβλητών συνεπάγεται σημαντική μείωση του R^2.

2. Η μέθοδος επιλογής προς τα μπρος: Σε αυτήν την μέθοδο ξεκινάμε με το μοντέλο που δεν έχει καμία μεταβλητή και στη συνέχεια προσθέτουμε κάθε φορά από μια μεταβλητή που έχει σημαντική συνεισφορά στο μοντέλο. Η πρώτη μεταβλητή είναι αυτή που έχει την υψηλότερη συσχέτιση με την εξαρτημένη μεταβλητή και στη συνέχεια επιλέγονται κατά σειρά σημαντικότητας οι επόμενες μεταβλητές.

3. Η διαδικασία της βηματικής παλινδρόμησης: Η διαδικασία της βηματικής παλινδρόμησης είναι παρόμοια με την μέθοδο επιλογής προς τα μπρος, με την μόνη διαφορά ότι σε κάθε βήμα ελέγχεται αν οι μεταβλητές οι οποίες έχουν ήδη προστεθεί είναι ακόμα σημαντικές.

Ωστόσο, ένα ζήτημα που δημιουργείται κατά την επιλογή των μεταβλητών, είναι και το **πρόβλημα της πολυσυγγραμμικότητας**. Εάν η τιμή μιας ανεξάρτητης μεταβλητής σχετίζεται με τις τιμές μιας ή περισσοτέρων άλλων ανεξάρτητων μεταβλητών τότε λέγεται ότι το μοντέλο παλινδρόμησης παρουσιάζει πολυσυγγραμμικότητα.

Ο έλεγχος πολυσυγγραμμικότητας μπορεί να γίνει με διάφορους τρόπους κι ένας από τους πιο δημοφιλής είναι ο υπολογισμός του Εκτιμητή Διόγκωσης της Διακύμανσης (Variance Inflation Factor – VIF). Ο εκτιμητής αυτός υπολογίζει κατά πόσο θα αυξηθεί η διακύμανση ενός εκτιμώμενου συντελεστή εάν η αντίστοιχη ανεξάρτητη μεταβλητή παρουσιάζει πολυσυγγραμμικότητα.

Στην πολλαπλή παλινδρόμηση είναι δυνατό κάποιες από τις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots, X_{p-1}$ να είναι γραμμικά εξαρτημένες με συνέπεια ο πίνακας πληροφορίας $X'X$ να μην αντιστρέφεται, αφού η ορίζουσά του είναι ίση με μηδέν και έτσι να μην μπορούν να βρεθούν οι εκτιμητές $\underline{b}$. Αυτό είναι γνωστό ως το πρόβλημα της πολυσυγγραμμικότητας ή πολλαπλής συγγραμμικότητας.

Στην περίπτωση που υπάρχουν σφάλματα στρογγύλευσης, μπορεί η ορίζουσα του πίνακα $X'X$ να μην είναι ακριβώς μηδέν, αλλά πολύ κοντά στο μηδέν δημιουργώντας ξανά το πρόβλημα στην αντιστροφή του πίνακα $X'X$. Αυτό αφορά το πρόβλημα της ασθενούς πολυσυγγραμμικότητας.

Το πρόβλημα της πολυσυγγραμμικότητας ή της ασθενούς πολυσυγγραμμικότητας μπορεί να προκύψει και από τις διαφορετικές μονάδες μέτρησης των μεταβλητών και έχει ως συνέπεια ορισμένες μεταβλητές να φαίνεται σημαντικές μέσω της p-τιμής σε κάποιο μοντέλο, ενώ παύουν να είναι σημαντικές όταν στο μοντέλο προσθέσουμε κι άλλες μεταβλητές. Αυτό μπορεί να δικαιολογηθεί αν σκεφτούμε πως η μεταβλητή που προσθέσαμε και που δείχνει να επηρεάζει την Y, δεν καταφέρνει να την ερμηνεύσει όσο μια άλλη μεταβλητή που δεν έχει ακόμη συμπεριληφθεί στο μοντέλο. Εκτός αυτού, η ύπαρξη της πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών αυξάνει το μέγεθος των τυπικών σφαλμάτων, με αποτέλεσμα να είναι πολύ μεγαλύτερα και τα διαστήματα εμπιστοσύνης.

ΚΕΦΑΛΑΙΟ 3

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

3.1 Εισαγωγη

Το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο στο οποίο όμως τα σφάλματα δεν ακολουθούν κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Η λογιστική παλινδρόμηση χρησιμοποιείται σε περιπτώσεις στις οποίες επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή (Y) είναι δίτιμη (δηλαδή παίρνει την τιμή 0 όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1 όταν υπάρχει το χαρακτηριστικό).

Η λογιστική παλινδρόμηση επινοήθηκε ως εναλλακτική επιλογή της γραμμικής διακριτικής ανάλυσης για την ταξινόμηση των στοιχείων, ονομαστικών ή τακτικών της εξαρτημένης μεταβλητής, με ευρεία απήχηση σε πολλά διαφορετικά επιστημονικά πεδία και κυρίως στην ιατρική και τις κοινωνικές επιστήμες. Χαρακτηριστικά, χρησιμοποιείται στην πρόβλεψή της:

- ✓ εμφάνισης ή μη μιας νόσου από ένα σύνολο διαφορετικών χαρακτηριστικών του πάσχοντος ατόμου, όπως η ηλικία, το φύλο κ.α.
- ✓ επιλογής ενός πολιτικού κόμματος με βάση την καταγραφή των δημογραφικών στοιχείων των πολιτών, όπως είναι η ηλικία, το φύλο, η φυλή, ο τόπος διαμονής, το εισόδημα και η προηγούμενη ψηφοφορία.
- ✓ πιθανότητα αποτυχίας μιας διεργασίας παραγωγής προϊόντος σε εργοστάσιο τροφίμων.
- ✓ πρόβλεψη της πρόθεσης αγοράς ενός αγαθού από έναν καταναλωτή (έρευνα αγοράς).
- ✓ πιθανότητα αθέτησης από δανειολήπτη της αποπληρωμής του δανείου του.

Όπως είπαμε και πιο πάνω η λογιστική παλινδρόμηση είναι η γενίκευση της απλής γραμμικής παλινδρόμησης, οπότε θα ξεκινήσουμε από το απλό γραμμικό μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ με i=1,2,...,n,}$$

όπου η Y_i είναι δυαδική, δηλαδή παίρνει ή την τιμή 0 ή την 1.

Μπορεί τα σφάλματα να μην κατανέμονται κανονικά, ωστόσο η μέση τιμή τους είναι μηδενική, δηλαδή $E(\varepsilon_i) = 0$ και έτσι προκύπτει ότι:

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = E(\beta_0 + \beta_1 X_i) + E(\varepsilon_i) = \beta_0 + \beta_1 X_i . \quad (1)$$

Επίσης, αφού η Y_i είναι μια δίτιμη μεταβλητή θα είναι μια μεταβλητή **Bernoulli**, οπότε ορίζουμε τις πιθανότητες ως εξής:

• Όταν το $Y_i = 1$ έχουμε $P(Y_i = 1) = \pi_i$

• Όταν το $Y_i = 0$ έχουμε $P(Y_i = 0) = 1 - \pi_i$, με π_i να είναι η πιθανότητα επιτυχίας.

Από τον ορισμό της μέσης τιμής εξασφαλίζουμε ότι:

$$E(Y_i) = 1 * \pi_i + 0 * (1 - \pi_i) = \pi_i . \quad (2)$$

Εξισώνοντας τις (1) και (2) βρίσκουμε:

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i \text{ και} \quad (3)$$

$$Var(Y_i) = \pi_i (1 - \pi_i).$$

Άρα ο αποκρινόμενος μέσος $E(Y_i)$ δηλώνει την πιθανότητα ότι η Y_i παίρνει την τιμή 1 όταν η προβλέπουσα μεταβλητή $X = X_i$.

Όταν η μεταβλητή Y είναι δίτιμη, δεν μπορεί να χρησιμοποιηθεί το γραμμικό μοντέλο και αυτό γιατί πέραν του γεγονότος ότι τα σφάλματα δεν είναι κανονικά κατανεμημένα, έχουν άνισες διασπορές.

Επιπλέον, ένα ακόμη πρόβλημα εντοπίζεται καθώς όπως έχουμε αναφέρει η εξαρτημένη μεταβλητή Y είναι δυαδική και παίρνει τις τιμές 0 και 1 με αποτέλεσμα να υπάρχει ο παρακάτω περιορισμός:

$$0 \le E(Y) = \pi \le 1.$$

Το πρόβλημα με τις άνισες διασπορές αντιμετωπίζεται χρησιμοποιώντας σταθμισμένα ελάχιστα τετράγωνα, ενώ με το να πάρουμε μεγάλο μέγεθος δείγματος, η μέθοδος ελαχίστων τετραγώνων παρέχει εκτιμητές που είναι ασυμπτωτικά κανονικοί ακόμα και όταν τα σφάλματα δεν ακολουθούν την κανονική κατανομή. Ο περιορισμός όμως στην τιμή της συνάρτησης απόκρισης είναι και το σημαντικότερο πρόβλημα και ο λόγος που η συνάρτηση απόκρισης είναι μη γραμμική.

3.2 Μελετη Του Απλου Λογιστικου Μοντελου

Το μοντέλο που χρησιμοποιούμε όταν η Y_i είναι δίτιμη είναι το λογιστικό, το οποίο ορίζεται ως εξής:

$$Y_i = E(Y_i) + \varepsilon_i,$$

όπου Y_i είναι ανεξάρτητη τυχαία μεταβλητή Bernoulli και:

$$E(Y_i) = \pi_i = \frac{e^{(\beta_0+\beta_1 X_i)}}{1+e^{(\beta_0+\beta_1 X_i)}} = \left[\frac{1+e^{(\beta_0+\beta_1 X_i)}}{e^{(\beta_0+\beta_1 X_i)}}\right]^{-1} = \left[\frac{1}{e^{(\beta_0+\beta_1 X_i)}}+1\right]^{-1} = \left[1+e^{(-\beta_0-\beta_1 X_i)}\right]^{-1}. \quad (4)$$

Η μορφή αυτή της αναμενόμενης συνάρτησης προκύπτει από την σχέση (3), δηλαδή $E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$. Κάνοντας την διαπίστωση ότι η συνάρτηση αυτή λαμβάνει τιμές από $-\infty$ έως $+\infty$ ενώ η αναμενόμενη συνάρτηση πρέπει να παίρνει τιμές στο διάστημα [0,1], προβαίνουμε στην χρήση των παρακάτω μετασχηματισμών. Ο πρώτος μετασχηματισμός επιτρέπει τη μετατροπή του ποσοστού σε ένα μέτρο, η τιμή του οποίου κυμαίνεται από 0 έως $+\infty$ και όχι από 0 έως 1 όπως συμβαίνει με το ποσοστό. Ο μετασχηματισμός αυτός επιτυγχάνεται λαμβάνοντας το λόγο συμπληρωματικών πιθανοτήτων (odds) του ποσοστού $\frac{\pi_i}{1-\pi_i}$ και όχι το ποσοστό αυτό καθεαυτό. Το odds μιας πιθανότητας είναι ο λόγος των συμπληρωματικών πιθανοτήτων, οπότε εάν η πιθανότητα εμφάνισης ενός ενδεχομένου συμβολιστεί με π_i και η πιθανότητα μη εμφάνισης με $(1-\pi_i)$, τότε το odds υπέρ του ενδεχομένου είναι $\frac{\pi_i}{1-\pi_i}$ ως προς ένα. Όταν η τιμή του ποσοστού πλησιάζει το μηδέν, τότε η τιμή της ποσότητας σχεδόν ταυτίζεται με την τιμή του ποσοστού, ενώ όταν η τιμή του ποσοστού πλησιάζει το ένα, τότε ο παρονομαστής της ποσότητας $\frac{\pi_i}{1-\pi_i}$ πλησιάζει το μηδέν και η τιμή της ποσότητας $\frac{\pi_i}{1-\pi_i}$ προσεγγίζει το $+\infty$. Ο δεύτερος μετασχηματισμός μετατρέπει το λόγο των συμπληρωματικών πιθανοτήτων του ποσοστού σε ένα μέτρο, η τιμή του οποίου κυμαίνεται από 0 έως 1. Στον μετασχηματισμό αυτό λαμβάνεται ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων. Το μέτρο που προκύπτει έπειτα από τους δύο αυτούς μετασχηματισμούς είναι το $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ και ονομάζεται logit μετασχηματισμός της πιθανότητας π_i.

Επομένως το logit είναι ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων. Ο μετασχηματισμός που περιλαμβάνει τα δύο παραπάνω βήματα ονομάζεται λογιστικός μετασχηματισμός.

Τελικά θα έχουμε ότι:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i \tag{5}$$

$$\Leftrightarrow \frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_i}$$

$$\Leftrightarrow \pi_i = e^{\beta_0+\beta_1 X_i}\left(1-\pi_i\right)$$

$$\Leftrightarrow \pi_i + \pi_i e^{\beta_0+\beta_1 X_i} = e^{\beta_0+\beta_1 X_i}$$

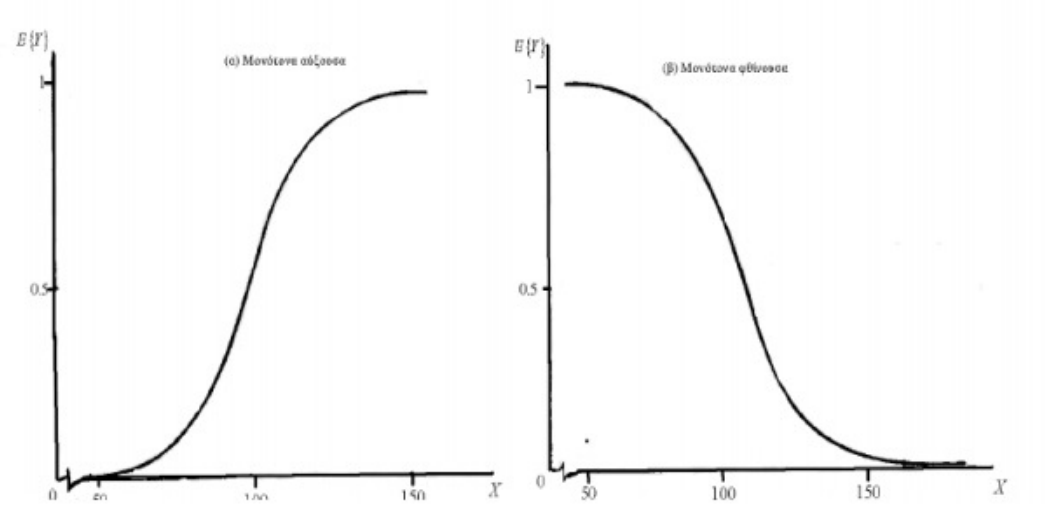
$$\Leftrightarrow \pi_i\left(1+e^{\beta_0+\beta_1 X_i}\right) = e^{\beta_0+\beta_1 X_i}$$

$$\Leftrightarrow \pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

$$\Leftrightarrow E\left(Y_i\right) = \frac{e^{(\beta_0+\beta_1 X_i)}}{1+e^{(\beta_0+\beta_1 X_i)}}.$$

Η αναμενόμενη λογιστική συνάρτηση είναι:

i. Είτε μονότονα αύξουσα συνάρτηση είτε μονότονα φθίνουσα,
ii. Είναι σχεδόν γραμμική στην περιοχή [0.2, 0.8],
iii. Πλησιάζει το 0 και 1 στις ακραίες τιμές της εμβέλειας του X όπως βλέπουμε και στην εικόνα 1.



Εικόνα 1: Παράδειγμα για τη λογιστική αναμενόμενη συνάρτηση, (α) μονότονα αύξουσα και (β) μονότονα φθίνουσα

3.3 Απλη Λογιστικη Παλινδρομηση

3.3.1 Εκτιμηση Παραμετρων Με Τη Μεθοδο Μεγιστης Πιθανοφανειας

Εξ' αιτίας της δυαδικής φύσης της εξαρτημένης μεταβλητής στο μοντέλο της λογιστικής παλινδρόμησης δεν είναι εφικτό να εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων, όπως εύκολα θα κάναμε με τη γραμμική παλινδρόμηση. Η δυσκολία αυτή μπορεί να ξεπεραστεί και η προσαρμογή του μοντέλου στα δεδομένα να πραγματοποιηθεί με τη μέθοδο της μέγιστης πιθανοφάνειας. Η μέθοδος της μέγιστης πιθανοφάνειας επιλέγει το σύνολο των τιμών των παραμέτρων του μοντέλου που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας.

Αφού τα Y_i είναι τυχαίες μεταβλητές Bernoulli όπου $\mathrm{P}(Y_i = 1) = \pi_i$ και $\mathrm{P}(Y_i = 0) = 1 - \pi_i$ η συνάρτηση πυκνότητας πιθανότητας είναι:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad Y_i = 0,1 \quad \text{και} \quad i = 1,\ldots,n \qquad (6)$$

Οι παρατηρήσεις Y_i είναι ανεξάρτητες οπότε η από κοινού συνάρτησης πιθανότητας θα είναι:

$$g(Y_1,\ldots\ldots,Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad (7)$$

Βέβαια είναι πιο εύκολο να δουλέψουμε με το λογάριθμο της από κοινού συνάρτησης και άρα η σχέση (7) θα γίνει:

$$\ln g(Y_1,\ldots\ldots,Y_n) = \ln \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$= \sum_{i=1}^{n}\left[Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i) \qquad (8)$$

Όμως λόγω των σχέσεων (4) και (5) μπορούμε να αντικαταστήσουμε το $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ και το $1-\pi_i$ οπότε θα έχουμε τη **λογαριθμική συνάρτηση πιθανοφάνειας** των εκτιμώμενων παραμέτρων:

$$\ln L(\beta_0,\beta_1)=\sum_{i=1}^{n}Y_i(\beta_0+\beta_1X_i)+\sum_{i=1}^{n}\ln\left(1-\frac{e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}}\right)$$

$$=\sum_{i=1}^{n}Y_i(\beta_0+\beta_1X_i)+\sum_{i=1}^{n}\ln\left(\frac{1+e^{\beta_0+\beta_1X_i}-e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}}\right)$$

$$=\sum_{i=1}^{n}Y_i(\beta_0+\beta_1X_i)+\sum_{i=1}^{n}\ln\left(1+e^{\beta_0+\beta_1X_i}\right)^{-1}$$

$$=\sum_{i=1}^{n}Y_i(\beta_0+\beta_1X_i)-\sum_{i=1}^{n}\ln\left(1+e^{\beta_0+\beta_1X_i}\right) \qquad (9)$$

Εδώ να αναφέρουμε πως δεν μπορούν να βρεθούν οι εκτιμήσεις των συντελεστών β_0 και β_1 από την μέθοδο μέγιστης πιθανοφάνειας, γιατί δεν υπάρχουν λύσεις κλειστής μορφής των τιμών των συντελεστών που να μεγιστοποιούν τη σχέση (9). Αφού βρεθούν οι εκτιμητές παλινδρομήσεως b_0 και b_1 μέσω επαναληπτικών αριθμητικών μεθόδων, αντικαθιστούμε τις τιμές τους στη σχέση (4) και έχουμε την προσαρμοσμένη λογιστική συνάρτηση απόκρισης:

$$\hat{\pi}=\frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}} \qquad (10)$$

Χρησιμοποιώντας το μετασχηματισμό logit θα έχουμε $\hat{\pi}'=\ln\frac{\hat{\pi}}{1-\hat{\pi}}$ και οπότε προκύπτει η προσαρμοσμένη αναμενόμενη λογιστική συνάρτηση (fitted logit) $\hat{\pi}'=b_0+b_1X$ (11).

3.3.2 Ερμηνεια Των Συντελεστων Παλινδρομησης

Ιδιαίτερα σημαντική είναι η ερμηνεία των συντελεστών της λογιστικής παλινδρόμησης. Όταν η εξαρτημένη μεταβλητή είναι ενδεικτική με Y=1 αν έχουμε την εμφάνιση ενός χαρακτηριστικού και Y=0 αν απουσιάζει η εμφάνιση του ίδιου χαρακτηριστικού, τότε ο συντελεστής b_1 ισούται με τον λόγο των logits

αυτών που έχουν το χαρακτηριστικό σε σχέση με αυτούς που δεν το έχουν, δηλαδή:

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right)-\ln\left(\frac{\pi_0}{1-\pi_0}\right)=\ln\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right)=\ln\left(\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}\right)=\ln\left(e^{\beta_1}\right)=b_1.$$

Έτσι στη λογιστική παλινδρόμηση, ο αντιλογάριθμος e^{b_1} του συντελεστή παλινδρόμησης b_1 μιας ενδεικτικής ανεξάρτητης μεταβλητής X_1, αποτελεί εκτίμηση του λόγου των odds αυτών που έχουν κάποιο χαρακτηριστικό σε σχέση με αυτούς που δεν το εμφανίζουν και ονομάζεται λόγος των εκτιμώμενων πιθανοτήτων:

$$e^{b_1}=e^{\ln\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right)}=\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right)=\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}.$$

Αν το b_1 είναι θετικό, ο παράγοντας e^{b_1} είναι μεγαλύτερος της μονάδας και έτσι ο εκτιμώμενος λόγος πιθανοτήτων αυξάνεται. Ανάλογα αν το b_1 είναι αρνητικό, ο παράγοντας e^{b_1} είναι μικρότερος της μονάδας και ο λόγος πιθανοτήτων μειώνεται.

Οι παράμετροι της λογιστικής παλινδρόμησης μπορούν να εκφραστούν και μέσα από το σχετικό λόγο των συμπληρωματικών πιθανοτήτων, δηλαδή το λόγο των odds που ονομάζεται odds ratio. Ο λόγος των odds ενός ατόμου με τιμές συμμεταβλητών X_1 σε σχέση με ένα άτομο με τιμές X_2 των ίδιων συμμεταβλητών δίνεται από την παρακάτω σχέση:

$$\text{Odds ratio}=\frac{\frac{\hat{\mathrm{p}}_1}{1-\hat{\mathrm{p}}_1}}{\frac{\hat{\mathrm{p}}_2}{1-\hat{\mathrm{p}}_2}}=\mathrm{e}^{(\mathbf{x}_1-\mathbf{x}_2)'\hat{\boldsymbol{\beta}}}$$

Αν το odds ratio είναι ίσο με τη μονάδα σημαίνει ότι τα odds των δύο ομάδων είναι ίσα. Πιο γενικά το odds ratio μας δείχνει πόσες φορές μεγαλύτερο ή μικρότερο είναι το ένα odds από το άλλο.

ΚΕΦΑΛΑΙΟ 4

ΠΟΛΛΑΠΛΗ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

4.1 Εισαγωγη

Πολλές φορές έχουμε περισσότερες από μια ανεξάρτητη μεταβλητή οπότε θα πρέπει να προεκτείνουμε το απλό λογιστικό μοντέλο σε πολλαπλό μοντέλο.

Το πολλαπλό λογιστικό μοντέλο είναι:

$$Y_i = E(Y_i) + \varepsilon_i = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_{p-1}X_{p-1}}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_{p-1}X_{p-1}}} + \varepsilon_i \qquad (1)$$

για να διευκολυνθούμε με τις πράξεις αλλά και για την απλοποίηση των σχέσεων θα χρησιμοποιήσουμε πίνακες και διανύσματα:

$$\underset{px1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \underset{px1}{X} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{pmatrix} \text{ και } \underset{px1}{X_i} = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{i,p-1} \end{pmatrix}. \qquad (2)$$

οπότε θα έχουμε

$$\beta' X = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} \text{ και } \beta' X_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}.$$

Με τη βοήθεια των πιο πάνω και με τα Yi να είναι ανεξάρτητες μεταβλητές Bernoulli μπορούμε να γράψουμε την αναμενόμενη λογιστική ως εξής:

Εναλλακτικά ισχύει ότι:

$$E(Y_i) = \pi_i = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} \qquad (3)$$

Παρατηρήσεις:

1. Πρέπει να πούμε εδώ ότι όλες οι σχέσεις που είχαμε δει στο απλό λογιστικό μοντέλο μπορούν να επεκταθούν και στο πολλαπλό λογιστικό μοντέλο
2. Επίσης να πούμε ότι και εδώ η αναμενόμενη λογιστική συνάρτηση είναι μονότονη και η καμπύλη έχει σχήμα s ή ανάποδο s σε σχέση με το $\beta' X$ και είναι σχεδόν γραμμική όταν $0.2 \le E(Y) \le 0.8$.
3. Οι μεταβλητές X μπορεί να είναι είτε ποσοτικές είτε ποιοτικές. Στην περίπτωση που είναι ποιοτικές αναπαριστώνται με δυαδικές μεταβλητές. Αν οι μεταβλητές είναι όλες ποιοτικές τότε το μοντέλο μπορεί να το ονομάσουμε και λογαριθμικό μοντέλο.

4.2 Εκτιμηση Παραμετρων

Όπως και στο απλό λογιστικό μοντέλο έτσι και εδώ η εκτίμηση παραμέτρων θα γίνει με τη βοήθεια της μεθόδου μέγιστης πιθανοφάνειας. Οπότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι:

$$\ln L(\beta) = \sum_{i=1}^{n} Y_i (\beta' X_i) - \sum_{i=1}^{n} \ln\left(1 + e^{\beta' X_i}\right) \qquad (4)$$

Για να βρούμε τους εκτιμητές θα πρέπει και πάλι να χρησιμοποιήσουμε κάποια αριθμητική μέθοδο η οποία θα μας δίνει τις τιμές των β_0, β_1, β_2,…, β_{p-1} που θα μεγιστοποιούν τη σχέση (4). Τις τιμές αυτές θα τις συμβολίζουμε με το

μοναδιαίο διάνυσμα $\underset{p\times 1}{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$.

Οπότε η προσαρμοσμένη λογιστική συνάρτηση απόκρισης θα έχει την μορφή:

$$\hat{\pi} = \frac{e^{b'X}}{1 + e^{b'X}} = \left[1 + e^{-b'X}\right]^{-1} \qquad (5)$$

Παρατηρήσεις:

1. Πιο πριν είπαμε πως οι εκτιμητές βρίσκονται με τη βοήθεια κάποιας αριθμητικής μεθόδου, μερικές φορές μπορεί να παρουσιαστεί κάποιο πρόβλημα σύγκλισης. Αυτό μπορεί να συμβεί όταν οι μεταβλητές πρόβλεψης είναι πάρα πολλές ή όταν κάποιες έχουν μεγάλη συσχέτιση. Αν συμβεί κάτι τέτοιο πρέπει να μειώσουμε τις μεταβλητές πρόβλεψης
2. Οι εκτιμητές μέγιστης πιθανοφάνειας μπορούν να βρεθούν και με τη μέθοδο των επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων
3. Όταν η λογιστική συνάρτηση δεν είναι μονότονη ή δεν έχει τη μορφή s θα πρέπει όλες τις μεταβλητές πρόβλεψης να τις μετατρέψουμε σε κατηγορικές, οπότε θα χρησιμοποιήσουμε το λογαριθμικό μοντέλο.

4.3 Ερμηνεια Των Συντελεστων Παλινδρομησης

Όσον αφορά τις ερμηνείες των συντελεστών παλινδρομήσεως, ο συντελεστής β_0 αποτελεί το ύψος της κλίσης της γραμμής παλινδρόμησης, ενώ καθένας από τους συντελεστές β_i εκφράζει το μέγεθος της συνεισφοράς της αντίστοιχης μεταβλητής.

- Θετική τιμή του συντελεστή δηλώνει ότι η επεξηγηματική μεταβλητή αυξάνει την πιθανότητα της επιτυχημένης έκβασης ενός γεγονότος, ενώ αρνητική τιμή σημαίνει ότι η μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης.
- Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επηρεάζει πολύ ισχυρά την πιθανότητα να συμβεί το γεγονός ή όχι, ενώ χαμηλή τιμή δηλώνει μικρή επίδραση της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.

Ισχύει και πάλι πως e^{b_k} είναι ο εκτιμώμενος λόγος πιθανοτήτων για την μεταβλητή X_k , υπό την προϋπόθεση ότι οι άλλες μεταβλητές πρόβλεψης είναι σταθερές και ότι το πολλαπλό μοντέλο παλινδρόμησης είναι πρώτης τάξεως και δεν περιέχει τετραγωνικούς ή μεγαλύτερου βαθμού όρους για τις μεταβλητές

πρόβλεψης. Όμως, όταν το πολλαπλό μοντέλο λογιστικής παλινδρόμησης δεν είναι ένα πρώτης τάξεως αλλά περιέχει τετραγωνικούς ή μεγαλύτερου βαθμού όρους για τις μεταβλητές πρόβλεψης, οι εκτιμώμενοι συντελεστές παλινδρόμησης δεν έχουν πλέον μια απλή ερμηνεία.

ΚΕΦΑΛΑΙΟ 5

ΠΡΟΓΡΑΜΜΑ ΣΤΑΤΙΣΤΙΚΗΣ SPSS

Στο κεφάλαιο αυτό θα αναπτύξουμε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης που στηρίζεται στη συνάρτηση παραγωγής Cobb-Douglas.

Πολλές φορές η εξίσωση που περιγράφει το υπό μελέτη μέγεθος δεν είναι γραμμική. Η συνάρτηση Cobb-Douglas που εμπλέκει το παραγόμενο προϊόν με την εργασία και το κεφάλαιο στην στοχαστική της μορφή έχει ως ακολούθως.

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} \varepsilon^{u_i} \qquad (6)$$

όπου Y είναι το προϊόν, X_2 είναι η εργασία που καταβάλλεται για την παραγωγή του προϊόντος αυτού, X_3 είναι το κεφάλαιο, u είναι ο διαταρακτικός όρος και e η βάση των φυσικών λογαρίθμων.

Με τη χρήση του λογαριθμικού μετασχηματισμού η εξίσωση από μη γραμμική γίνεται γραμμική

$$lnY_i = ln\beta_1 + \beta_2 lnX_{2i} + \beta_3 lnX_{3i} + u_i = \beta_0 + \beta_2 lnX_{2i} + \beta_3 lnX_{3i} + u_i \qquad (7)$$

όπου $\beta_0 = ln\beta_1$.

Βλέπουμε ότι η συνάρτηση είναι μη γραμμική στις μεταβλητές Y και X αλλά γραμμική στους λογαρίθμους αυτών των μεταβλητών. Η εξίσωση (7) λέγεται log-linear μοντέλο.

Οι ιδιότητες της συνάρτησης αυτή είναι:

1. β_2 είναι η μερική ελαστικότητα της αξίας του εξαγόμενου προϊόντος αναφορικά με την εργασία. Δηλαδή μετράει την ποσοστιαία μεταβολή της αξίας του προϊόντος για 1% μεταβολή της εργασίας κρατώντας το κεφάλαιο σταθερό.
2. Παρόμοια, η β_3 είναι η ελαστικότητα όσον αφορά το κεφάλαιο.
3. Το άθροισμα (β1+β2) δίνει πληροφορία σχετικά με την απόδοση κλίμακας ή επιστροφή στην κλίμακα το οποίο είναι η μεταβολή της παραγωγής αναφορικά με αλλαγές σε εργασία και κεφάλαιο. Αν ισούται με 1 τότε σταθερές αποδόσεις κλίμακας. Δηλαδή αν διπλασιάσουμε τις εισαγωγές (εργασία, κεφάλαιο) θα διπλασιάσουμε την παραγωγή. Αν το άθροισμα είναι μικρότερο από 1 έχουμε μειούμενες αποδόσεις κλίμακας και ο διπλασιασμός πόρων θα αυξήσεις λιγότερο από το διπλάσιο την παραγωγή. Αντίστοιχα για άθροισμα μεγαλύτερο του 1 έχουμε αυξανόμενες αποδόσεις και ο διπλασιασμός πόρων θα υπερδιπλασιάσει την παραγωγή

Για την ανάπτυξη του μοντέλου Cobb-Douglas μέσω της 7 πήραμε δεδομένα από την ετήσια έκθεση 2005. Annual Survey of Manufacturers. Sector31: Supplemental Statistics for U.S.

State	Product.Output	Labor.Input	Capital.Input
Alabama	38372840	424471	2689076
Alasca	1805427	19895	57997
Arizona	23736129	206893	2308272
Arkansas	26981983	304055	1376235
California	217546032	1809756	13554116
Colorado	19462751	180366	1790751
Connecticut	28972772	224267	1210229
Delaware	14313157	54455	421064
District of Columbia	159921	2029	7188
Florida	47289846	471211	2761281
Georgia	63015125	659379	3540475
Havai	1809052	17528	146371
Idaho	10511786	75414	848220
Illinois	105324866	963156	5870409
Indiana	90120459	835083	5832503
Iowa	39079550	336159	1795976
Kansas	22826760	246144	1595118
Kentucky	38686340	384484	2503693
Louisiana	69910555	216149	4726625
Maine	7856947	82021	415131
Maryland	21352966	174855	1729116
Massachusetts	46044292	355701	2706065
Michigan	92335528	943298	5294356
Minnesota	48304274	456553	2833525
Mississipi	17207903	267806	1212281
Missouri	47340157	439427	2404122
Montana	2644567	24167	334008
Nebraska	14650080	163637	627806
Nevada	7290360	59737	522335
New Hampshire	9188322	96106	507488
New Jersey	51298516	407076	3295056
New Mexico	20401410	43079	404749
New York	87756129	727177	4260353
North Carolina	101268432	820013	4086558
North Dakota	3556025	34723	184700
Ohio	124986166	1174540	6301421
Oklahoma	20451196	201284	1327353
Oregon	34808109	257820	1456683
Pensilvania	104858322	944998	5896392
Rhode Island	6541356	68987	297618
South Carolina	37668126	400317	2500071
South Dakota	4988905	56524	311251
Tennessee	62828100	582241	4126465

Texas	172960157	1120382	11588283
Utha	15702637	150030	762671
Vermont	5418786	48134	276293
Virginia	49166991	425346	2731669
Washington	46164427	313279	1945860
West Virginia	9185967	89639	685587
Wisconsin	66964978	694628	3902823
Wyoming	2979475	15221	361536

Υποθέτοντας ότι το μοντέλο 7 ικανοποιεί τις προϋποθέσεις της γραμμικής παλινδρόμησης αναπτύσσουμε την ακόλουθη παλινδρόμηση:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,982[a]	,964	,963	,26675

a. Predictors: (Constant), log.Capital, log.Labor

b. Dependent Variable: log.Product

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91,925	2	45,962	645,931	<,001[b]
	Residual	3,416	48	,071		
	Total	95,340	50			

a. Dependent Variable: log.Product

b. Predictors: (Constant), log.Capital, log.Labor

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,888	,396		9,812	<,001
	log.Labor	,468	,099	,464	4,734	<,001
	log.Capital	,521	,097	,528	5,380	<,001

a. Dependent Variable: log.Product

Οι παραπάνω πίνακες μεταφράζονται στην ακόλουθη εξίσωση:

$$lnY_i = 3.888 + 0.468lnX_{2i} + 0.52lnX_{3i}$$

Βλέπουμε ότι οι ελαστικότητες εργασίας και κεφαλαίου είναι 0,468 και 0,521 αντίστοιχα. Αν δηλαδή αυξηθεί η εργασία κατά 1% (κρατώντας το κεφάλαιο σταθερό) η παραγωγή θα αυξηθεί κατά 0,47%. Αντίστοιχα ισχύουν και για το κεφάλαιο για το οποίο μεταβολή κατά 1% θα επιφέρει αύξηση στην παραγωγή 0,52%. Αν προσθέσουμε τους συντελεστές β_1 και β_2 θα πάρουμε 0,99 το οποίο μας δίνει και την απόδοση κλίμακας την οποία θα χαρακτηρίζαμε σαν σταθερή.

Από στατιστικής άποψης το R^2 = 0,964 μας πληροφορεί ότι η μεταβολή στην παραγωγή (στον λογάριθμό της) οφείλεται κατά 96% στην μεταβολή εργασίας και κεφαλαίου. Επίσης το ότι η τιμή της p-value για την anova είναι μικρότερη από 0,05 (p-value=<0.01 από πίνακα) μας πληροφορεί ότι το μοντέλο είναι σημαντικό. Δηλαδή έχει νόημα. Τέλος από τον πίνακα των συντελεστών βλέπουμε ότι και οι δύο μεταβλητές (εργασία και κεφάλαιο) έχουν p-value<0,05 και άρα είναι στατιστικά σημαντικές. Το ίδιο ισχύει και για το σταθερό όρο ο οποίος είναι μεγαλύτερος του μηδενός και ο οποίος μας λέει ότι δεν μπορεί να μηδενιστεί η παραγωγή. Πρέπει ο λογάριθμός της να ισούται με τουλάχιστον 3,88 (ίσως γιατί διαφορετικά δε θα μπορούσε να ζήσει ο άνθρωπος).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία επιχειρήθηκε η αποτύπωση των τεχνικών ανάλυσης της παλινδρόμησης προκειμένου να γίνουν αντιληπτά τα διάφορα μοντέλα της τα οποία είναι θεμελιωμένα σε ερευνητικές υποθέσεις που επικυρώνονται ή όχι από τα δεδομένα της πραγματικότητας. Από τη παρουσίαση των τεχνικών ανάλυσης της παλινδρόμησης προέκυψαν χρήσιμα συμπεράσματα καθώς και ευρύτερες σκέψεις, που συντελούν σε μια βαθύτερη και πιο ουσιαστική μελέτη των στατιστικών μοντέλων. Ειδικότερα μέσω της μελέτης των περισσότερο δημοφιλών μοντέλων παλινδρόμησης μπορέσαμε να αντλήσουμε γνώσεις και παραδείγματα που μας βοήθησαν στην προσέγγιση των ερευνητικών σκοπών της εργασίας μας. Τα μοντέλα που μελετήθηκαν ήταν της απλής γραμμικής παλινδρόμησης, της πολλαπλής γραμμικής παλινδρόμησης, του απλού λογιστικού μοντέλου, της πολλαπλής λογιστικής παλινδρόμησης, εφαρμογής των δεδομένων με χρήση του Προγράμματος Στατιστικής SPSS.

Στην πρώτη και πιο απλή περίπτωση παλινδρόμησης, της απλής γραμμικής, είδαμε την αξία της εφαρμογής της σε πειραματικές (ανεξάρτητη και εξαρτημένη μεταβλητή) και μη πειραματικές μελέτες. Είναι το πλέον διαδομένο από όλα τα μοντέλα, διαθέτοντας εκτεταμένες δυνατότητες ερευνητικών χρήσεων. Το επόμενο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης μας αποκάλυψε τις δυνατότητες εφαρμογής του σε επιχειρησιακά και εν γένει οικονομικά περιβάλλοντα με έμφαση στην εξέταση του επενδυτικού πλαισίου και του βαθμού διακινδύνευσης. Χρησιμοποιώντας δύο ή περισσότερες ανεξάρτητες

μεταβλητές είμαστε σε θέση να ερμηνεύσουμε με μεγάλο βαθμό ακριβείας ένα φυσικό φαινόμενο και να εξάγουμε ορθότερες προβλέψεις. Στη συνέχεια παρουσιάσαμε το λογιστικό μοντέλο ως ένα μη γραμμικό μοντέλο στο οποίο τα σφάλματα δεν ακολουθούν κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή χρησιμοποιείται δηλαδή σε περιπτώσεις στις οποίες επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού στοιχείου, ή ενός συμβάντος που έχει εκδηλωθεί. Έτσι οι περιπτώσεις συνδέονται με αυτό το μοντέλο αφορούν τον ιατρικό κλάδο, την πολιτική, την παραγωγή και τις τρέχουσες οικονομικές συνθήκες. Ακολούθως, όταν έχουμε περισσότερες από μια ανεξάρτητη μεταβλητή οπότε θα πρέπει να προεκτείνουμε το απλό λογιστικό μοντέλο σε πολλαπλό μοντέλο. Αξίζει να σημειωθεί ότι το μοντέλο της πολλαπλής λογιστικής παλινδρόμησης σε πολλές περιπτώσεις εφαρμογής του δεν έχει μονοσήμαντες ερμηνείες. Τέλος, παρουσιάσαμε το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης, στηριζόμενο στη συνάρτηση παραγωγής Cobb-Douglas.

Από την παρουσίαση των κυριότερων μοντέλων παλινδρόμησης μπορέσαμε να διεισδύσουμε στις πτυχές εφαρμογής τους και αντιληφθούμε την σημασία τους για την καθημερινότητά των ανθρώπινων κοινωνιών. Η χρήση τους μας απέδειξε την ανάγκη περαιτέρω διερεύνησης των δυνατοτήτων τους που θα αποσκοπεί στην επέκταση του πρακτικού τους αποτυπώματος στην ιατρική, την πολιτική, την οικονομία και άλλους τομείς της δημόσιας ζωής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Υπουργείο Παιδείας, Έρευνας και Θρησκευμάτων Ινστιτούτο Εκπαιδευτικής Πολιτικής: Αδαμόπουλος Λ., Δαμιανού Χ. , Σβέρκος Α., (1999), *ΜΑΘΗΜΑΤΙΚΑ ΚΑΙ ΣΤΟΙΧΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ Γ΄ ΓΕΝΙΚΟΥ ΛΥΚΕΙΟΥ*, Ινστιτούτο Τεχνολογίας Υπολογιστών και Εκδόσεων «ΔΙΟΦΑΝΤΟΣ», Αθήνα.
- Κιόχος Π., Κιόχος Α., (2010), *ΣΤΑΤΙΣΤΙΚΗ ΓΙΑ ΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΤΗΝ ΟΙΚΟΝΟΜΙΑ*, Εκδόσεις Ελένη Κιόχου, Αθήνα
- Χαλικιάς Μ., Λάλου Π., Μανωλέσου Α., (2015), *Μεθοδολογία έρευνας και εισαγωγή στη Στατιστική Ανάλυση Δεδομένων με το IBM SPSS STATISTICS*, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα
- Ηλιοπούλου Π., (2015), *ΓΕΩΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ*, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα
- Πετρίδης Δ., (2015), *ΑΝΑΛΥΣΗ ΠΟΛΥΜΕΤΑΒΛΗΤΩΝ ΤΕΧΝΙΚΩΝ, ΕΦΑΡΜΟΓΕΣ ΠΕΡΙΠΤΩΣΕΩΝ*, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα
- ΒΙΚΙΠΑΙΔΕΙΑ, η ελεύθερη εγκυκλοπαίδεια, Διαθέσιμο στην ηλεκτρονική διεύθυνση: *https://el.wikipedia.org*
- Gujarati D.*, Porter D. "Basic Econometrics", Mc Graw Hill Fifth Edition.*