ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΕΞΕΤΑΣΗ ΤΗΣ ΣΧΕΣΗΣ ΜΕΤΑΞΥ ΜΕΤΑΒΛΗΤΩΝ-ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΕΛΕΓΧΟΣ Χ ΤΕΤΡΑΓΩΝΟ»

ΝΤΟΣΚΑΣ ΔΗΜΗΤΡΙΟΣ

ΓΑΣΠΑΡΙΝΑΤΟΣ ΓΕΩΡΓΙΟΣ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ : ΜΕΓΑΡΙΤΗΣ ΑΘΑΝΑΣΙΟΣ

ΜΕΣΟΛΟΓΓΙ, 2017

ΕΠΙΣΥΝΑΨΗ

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, του Εκπαιδευτικού Προσωπικού του Τμήματος Λογιστικής και Χρηματοοικονομικής ή του Α.Τ.Ε.Ι. Δυτικής Ελλάδος.

ΝΤΟΣΚΑΣ ΔΗΜΗΤΡΙΟΣ

ΓΕΩΡΓΙΟΣ ΓΑΣΠΑΡΙΝΑΤΟΣ

ΠΡΟΛΟΓΟΣ

Η εργασία αυτή εκπονήθηκε κατά τη διάρκεια των σπουδών μας στο τμήμα Λογιστικής και Χρηματοοικονομικής, στη σχολή Διοίκησης και Οικονομίας, στο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδας.

Στο σημείο αυτό θα θέλαμε να εκφράσουμε θερμές ευχαριστίες στα πρόσωπα που μας στήριξαν καθ' όλη τη διάρκεια των σπουδών μας και ιδιαίτερα στον επιβλέποντα καθηγητή της παρούσας μελέτης κύριο Μεγαρίτη του οποίου η βοήθεια ήταν πολύτιμη για τη διεκπεραίωσή της.

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας μελέτης είναι η διερεύνηση της σχέσης μεταξύ δυο ή περισσότερων μεταβλητών. Κατά την απλή γραμμική παλινδρόμηση διερευνάται η σχέση που συνδέει δύο μεταβλητές ενώ κατά την πολλαπλή γραμμική παλινδρόμηση εξετάζεται η σχέση μεταξύ περισσότερων από δύο μεταβλητών. Ο έλεγχος Χ τετράγωνο χρησιμοποιείται για τον έλεγχο της σχέσης μεταξύ ποιοτικών μεταβλητών.

Η ανάλυση και επεξεργασία των δεδομένων που χρησιμοποιήθηκαν για τις ανάγκες της εργασίας, πραγματοποιήθηκε με την εκτεταμένη χρήση της στατιστικής για την εξαγωγή των στατιστικών συμπερασμάτων. Για το σκοπό αυτό χρησιμοποιήθηκε το στατιστικό πακέτο SPSS 20. Με βάση αυτό έγιναν οι απαιτούμενες στατιστικές αναλύσεις και έτσι εξήχθησαν αποτελέσματα για την απλή γραμμική παλινδρόμηση, την πολλαπλή παλινδρόμηση και τον έλεγχο X^2 με σκοπό τόσο τη διερεύνηση της σχέσης μεταξύ 2 ή περισσότερων μεταβλητών όσο και μεταξύ ποιοτικών μεταβλητών.

ΠΕΡΙΕΧΟΜΕΝΑ

I. Πρόλογος 2
II. Περίληψη 3
III. Εισαγωγή 6
Κεφάλαιο 1 :Στατιστική 7
1.1 Ιστορική αναδρομή της Στατιστικής 7
1.2 Βασικές Έννοιες Στατιστικής 8
Κεφάλαιο 2: Παλινδρόμηση 10
2.1 Εισαγωγή στην παλινδρόμηση 10
2.2 Απλή παλινδρόμηση 11
2.3 Πολλαπλή παλινδρόμηση 11
2.4 Γραμμική παλινδρόμηση 11
2.5 Μη γραμμική παλινδρόμηση 12
Κεφάλαιο 3: Απλή γραμμική παλινδρόμηση 15
3.1 Βασικές έννοιες 15
3.2 Διάγραμμα διασποράς 17
3.3 Μέθοδος ελαχίστων τετραγώνων 18
3.4 Συντελεστής γραμμικής συσχέτισης του Pearson 20
3.5 Συντελεστής προσδιορισμού 22
3.6 Απλή γραμμική παλινδρόμηση με χρήση του SPSS 24
Κεφάλαιο 4: Πολλαπλή γραμμική παλινδρόμηση 33
4.1 Βασικές έννοιες 33
4.2 Εξίσωση ελαχίστων τετραγώνων 36
4.3 Συντελεστής πολλαπλής συσχέτισης 38
4.4 Συντελεστής μερικής συσχέτισης 39
4.5 Συντελεστής πολλαπλού προσδιορισμού 40
4.6 Πολλαπλή γραμμική παλινδρόμηση με χρήση του SPSS 41
Κεφάλαιο 5: Έλεγχος ανεξαρτησίας ποιοτικών μεταβλητών 59

5.1 Στατιστικοί έλεγχοι υποθέσεων 59

5.2 Έλεγχος Χ-τετράγωνο..... 61

5.3 Εξέταση της σχέσης μεταξύ ποιοτικών μεταβλητών με χρήση του SPSS 64

Συμπεράσματα 81

Βιβλιογραφία 83

ΕΙΣΑΓΩΓΗ

Η στατιστική αποτελεί ένα κλάδο των εφαρμοσμένων μαθηματικών που έχει ως αντικείμενο την εξέταση και μελέτη φυσικών, κοινωνικών και οικονομικών φαινομένων. Εφαρμόζονται μέθοδοι συλλογής και επεξεργασίας δεδομένων και πληροφοριών που περιλαμβάνουν ομάδες έμψυχων ή άψυχων αντικειμένων. Με αυτόν τον τρόπο προκύπτουν συμπεράσματα τα οποία είναι χρήσιμα για τον ερευνητή και τον καθοδηγούν στην εξαγωγή αποτελεσμάτων και στη λήψη συγκεκριμένων αποφάσεων (Ιωαννίδης, 2005) (Ζαχαροπούλου, 2010) (Champkin, 2013).

Η στατιστική χρησιμοποιείται σχεδόν σε όλες τις επιστήμες και σε πολλούς τομείς της καθημερινής ζωής. Κατά κύριο λόγο χρησιμοποιείται στη Βιολογία, στην Ιατρική, στη Μετεωρολογία, στην Ψυχολογία, στην Κοινωνιολογία αλλά κατά βάση έχει ευρύτατη εφαρμογή στις Οικονομικές Επιστήμες. Η στατιστική επιστήμη συνδυαστικά με τα μαθηματικά και τη Μαθηματική Οικονομική δημιούργησε τον κλάδο της Οικονομετρίας. Η οικονομετρία βασιζόμενη στις στατιστικές μεθόδους αποσκοπεί στην εμπειρική εκτίμηση και επαλήθευση της οικονομικής θεωρίας (Ζαχαροπούλου, 2010).

Η στατιστική μέθοδος καλείται να διευκολύνει τον ερευνητή ώστε να μπορεί να ρυθμίσει τις συνθήκες του φαινομένου το οποίο εξετάζει. Αυτές οι συνθήκες μπορεί να είναι σταθερές ή να μεταβάλλονται ανάλογα με ενδογενείς ή εξωγενείς παράγοντες. Ο ερευνητής έχει τη δυνατότητα να μεταβάλλει τις συνθήκες και να μελετήσει την επίδραση που ασκούν στο φαινόμενο που εξετάζει (Ιωαννίδης, 2005) (Champkin, 2013).

ΚΕΦΑΛΑΙΟ 1 :ΣΤΑΤΙΣΤΙΚΗ

1.1 Ιστορική αναδρομή της Στατιστικής

Αρχικά η στατιστική και οι μέθοδοί της χρησιμοποιήθηκαν για τη μελέτη και την επίλυση των δημογραφικών προβλημάτων του πληθυσμού των πόλεων της Φρανκφούρτης το 1440 και της Νυρεμβέργης το 1526. Οι τεχνικές στατιστικής που χρησιμοποιήθηκαν σε πρώιμο στάδιο άρχιζαν να βελτιώνονται με τις στατιστικές γάμων, γεννήσεων ή θανάτων οι οποίες έγιναν υποχρεωτικές από τον 16^{ο} αιώνα. Ο πρώτος πίνακας θνησιμότητας συντάχθηκε το 1661 από τον Άγγλο John Graunt ο οποίος ήταν θεμελιωτής της στατιστικής δημογραφίας. Σημαντικός σταθμός για την εξέλιξη της επιστήμης της στατιστικής αποτέλεσε η εθνική απογραφή του πληθυσμού της Γαλλίας από τον Vauban στις αρχές του 18^{ου} αιώνα (Champkin, 2013).

Η Στατιστική ως έννοια έκανε την εμφάνισή της για πρώτη φορά από τους μυθικούς χρόνους όταν δημιουργήθηκαν για πρώτη φορά οι οργανωμένες κοινωνίες. Μια από τις πρώτες γραφές στατιστικού τύπου που πραγματοποιήθηκε από τον Όμηρο και περιελάμβανε στατιστικές πληροφορίες είναι ο νέων κατάλογος δηλαδή ο κατάλογος των πλοίων που ανήκαν στους Αχαιούς κατά τον Τρωικό πόλεμο[1]. Ο κατάλογος αυτός βοήθησε το έργο των ιστορικών στο να αποσπάσουν σημαντικές εκτιμήσεις όσον αφορά την οικονομική ευρωστία, τον αριθμό του πληθυσμού των πόλεων-κρατών που συμμετείχαν καθώς και σημαντικά στοιχεία για την τότε ναυπηγική, ναυτιλία και ναυτική τέχνη. Ενώ η πρώτη ιστορική συλλογή που πραγματοποιήθηκε και σχετίζεται καθαρά με στατιστικά στοιχεία θεωρείται η απογραφή πληθυσμού από τον Αυτοκράτορα της Κίνας Γιάο (Yao) το 2238 π.Χ. Παρόμοιες απογραφές με στατιστικά δεδομένα πραγματοποιήθηκαν και από άλλους αρχαίους λαούς. Ενδεικτικά αναφέρουμε τους Αιγύπτιους, τους Βαβυλώνιους, τους Πέρσες, τους αρχαίους Έλληνες και τους Ρωμαίους. Η πιο χαρακτηριστική από όλες είναι η απογραφή του Οκταβιανού Αυγούστου (Ζαχαροπούλου, 2010).

Κατά την αρχαιότητα ο πρώτος στόχος με τον οποίο πραγματοποιούταν η συλλογή των δεδομένων ήταν η συγκέντρωση και η φορολόγηση των πολιτών αλλά και των πόλεων. Επιπλέον χαρακτηριστικά στατιστικά στοιχεία αποτελούν ο μέδιμνος και ο ίππος στην Αρχαία Αθήνα κατά την Αρχαϊκή εποχή. Άλλα πολύ σημαντικά γεγονότα που στηρίχτηκαν σε στατιστικά στοιχεία αποτελέσαν οι φυλές και οι Δήμοι στην Εκκλησία στο Δήμο της Αθήνας καθώς και οι ψηφοφορίες που λάμβαναν χώρα. Αργότερα, οι Ρωμαίοι στηριζόμενοι σε στατιστικά στοιχεία προέβησαν στη διοικητική διαίρεση της Αυτοκρατορίας τους. Στη συνέχεια ακολούθησε και η Βυζαντινή Αυτοκρατορία δημιουργώντας τα βυζαντινά θέματα βάση στατιστικών δεδομένων (Κικιλίας, et al., 2001).

Η δημογραφική στατιστική που αναφέρεται στη συλλογή πληροφοριών του πληθυσμού και της οικονομίας μιας χώρας ή μιας πόλης ξεκίνησε τη περίοδο της Αναγέννησης και πιο συγκεκριμένα στις πόλεις της Ιταλίας, στη Βενετία και στη Φλωρεντία όπου γνώρισε γρήγορη επέκταση σε όλα τα τότε Βασίλεια της Ευρώπης. Μετά το τέλος του 11ου αιώνα, επί εποχής Γουλιέλμου του Κατακτητή, σημειώθηκε μια πολύ σπουδαία στατιστική απογραφή που περιελάμβανε πολλές μονάδες παραγωγής της Αγγλίας, όπως μεταλλεία, ιχθυοτροφεία κ.λπ. Οι γάμοι και οι θάνατοι

[1] Βλέπε Ιλιάδα β' (στ. 494-759)

έδωσαν γρήγορη ανάπτυξη στην στατιστική. Ειδικότερα, οι θάνατοι οι οποίοι είχαν αυξηθεί με μεγάλο ρυθμό και παρατηρήθηκαν αργότερα εξαιτίας του πολέμου, των επιδημιών και των λιμοκτονιών ώθησαν και συνέβαλλαν στην ανάπτυξη της στατιστικής έρευνας μέσα από την καταγραφή των αιτιών και των απωλειών. Κατά συνέπεια μετά το 1348 άρχισαν να γίνονται οι καταγραφές θανάτων από την πανώλη η οποία αποτέλεσε μια από τις φοβερότερες ασθένειες και διήρκησε τέσσερις αιώνες. Στις καταγραφές αυτές που σημειώθηκαν κατά την ίδια περίοδο περιελήφθησαν ακόμα περισσότερες απώλειες που προήλθαν από διάφορες αιτίες. Η πρώτη έρευνα που αφορούσε δείγμα πληθυσμού ξεκίνησε το 1620 από τον Άγγλο έμπορο Τζον Γκράουντ στο Λονδίνου σε ορισμένες οικογένειες. Από την έρευνα διαπίστωσε ότι ανά 88 άτομα υπήρχαν τρεις θάνατοι. Μέσα από αυτά τα στατιστικά στοιχεία καθώς και τις 13.200 απώλειες που καταγράφτηκαν διαπιστώθηκε ότι οι κάτοικοι που ζούσαν στο Λονδίνο το 1620 δεν ξεπερνούσε τους 387.000 κατοίκους (Ιωαννίδης, 2005).

Συνεπώς το έτος 1663 τέθηκε σαν αρχή της Στατιστικής Επιστήμης από πολλούς ερευνητές και επιστήμονες. Ειδικότερα το βιβλίο του John Graunt που εκδόθηκε το ίδιο έτος και είχε ως τίτλο «Φυσικές και Πολιτικές παρατηρήσεις της Θνησιμότητας» αποτέλεσε αφετηρία για την Στατιστική (Willcox, 1938). Στη συνέχεια από τον 16ο μέχρι τον 19ο αιώνα σημειώθηκε ραγδαία ανάπτυξη του εμπορίου η οποία οδήγησε στον εξαναγκασμό των αρχών των κρατών στη μελέτη των νέων οικονομικών δεδομένων του εμπορίου των μεταφορών και των βιομηχανιών καθώς και του εργατικού δυναμικού. Σήμερα η στατιστική έρευνα από μια απλή μαθηματική τεχνική έχει μετατραπεί και αναχθεί σε σπουδαία αυτοτελή επιστήμη ακολουθώντας και εφαρμόζοντας ιδιαίτερες μεθόδους στατιστικής ανάλυσης (Champkin, 2013).

1.2 Βασικές Έννοιες Στατιστικής

Ένας ερευνητής προκειμένου να καλύψει τους ερευνητικούς του στόχους συγκεντρώνει πληροφορίες, δεδομένα ή διαφορετικά παρατηρήσεις έτσι ώστε να εξετάσει ένα συγκεκριμένο φαινόμενο ή να πραγματοποιήσει μια συγκεκριμένη μελέτη. Αυτά τα δεδομένα ονομάζονται **στατιστικά δεδομένα.** Επιπλέον όλα αυτά που χρησιμοποιούνται για την πραγματοποίηση της έρευνας όπως είναι τα πρόσωπα, οι καταστάσεις, τα γεγονότα, οι έννοιες, οι ιδέες και τα αισθήματα από τα οποία αντλούνται όλα αυτά τα δεδομένα και οι πληροφορίες ονομάζονται **στατιστικές μονάδες.** Το γνώρισμα κάθε στατιστικής μονάδας που χρησιμοποιείται σε μια μελέτη ονομάζεται χαρακτηριστικό. Ο ορισμός του χαρακτηριστικού θα πρέπει να χρησιμοποιείται ορθά με σκοπό να αποφεύγονται συγχύσεις και οι τυχόν ανόμοιες συγκρίσεις μεταξύ των μεγεθών οι οποίες οδηγούν συχνά σε αντιρρήσεις όσον αφορά τα στατιστικά δεδομένα. Οι στατιστικές μονάδες που παρουσιάζουν ένα ή περισσότερα κοινά χαρακτηριστικά ονομάζονται στο σύνολό τους **στατιστικός πληθυσμός** ή απλά **πληθυσμός** (Ζαχαροπούλου, 2010).

Η μελέτη δεν μπορεί να περιλαμβάνει μεγάλο όγκο πληροφοριών και δεδομένων δηλαδή δεν είναι δυνατό να βασιστεί σε όλες τις στατιστικές μονάδες που περιλαμβάνει λόγω αδυναμίας συγκέντρωσης των απαιτούμενων στοιχείων και λόγω υψηλού κόστους. Συνεπώς η μελέτη περιορίζεται σε ένα μόνο μέρος του πληθυσμού το οποίο ονομάζεται **δείγμα** (Ιωαννίδης, 2005).

Το **μέγεθος του δείγματος** μας επισημαίνει τις στατιστικές μονάδες από τις οποίες αποτελείται ένα δείγμα. Αξίζει να σημειωθεί ότι οι στατιστικές μονάδες αποτελούνται από διάφορα χαρακτηριστικά. Όταν μια ομάδα στατιστικών μονάδων έχει ένα κοινό χαρακτηριστικό δίνεται η δυνατότητα να καταταχθεί στον ίδιο πληθυσμό και να ενταχθεί δηλαδή σε μια κατηγορία. Για παράδειγμα, όλοι οι Έλληνες έχουν ένα κοινό χαρακτηριστικό ότι κατάγονται από την Ελλάδα. Όλοι οι Ευρωπαίοι έχουν ένα κοινό χαρακτηριστικό ότι είναι από την Ευρώπη αλλά άλλοι είναι Έλληνες, άλλοι Γάλλοι και άλλοι Ιταλοί. Επομένως σε αυτόν τον πληθυσμό μπορούν να μελετηθούν και άλλα χαρακτηριστικά όπως εθνικότητα, περιοχή που κατοικεί, θρησκεία, φύλο, ηλικία κλπ. Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε ένα πληθυσμό ονομάζονται **μεταβλητές** οι οποίες συμβολίζονται με κεφαλαία γράμματα συνήθως χρησιμοποιείται το γράμμα Χ. Οι πιθανές τιμές που παίρνει μια μεταβλητή που εξετάζουμε ονομάζονται τιμές της μεταβλητής και τις συμβολίζουμε με ένα x_i.

Έτσι η ποικιλία και η ευρεία γκάμα των διαφορετικών χαρακτηριστικών των στατιστικών μονάδων τα κατατάσσει σε 2 κατηγορίες: τα **ποιοτικά** και τα **ποσοτικά** χαρακτηριστικά. Ποιοτικά χαρακτηριστικά ή ποιοτικές ή κατηγορικές μεταβλητές ονομάζονται οι μεταβλητές των οποίων οι τιμές μπορούν να ταξινομηθούν σε κατηγορίες και δεν εκφράζουν ποσότητα δηλαδή δεν εκφράζουν κάτι το οποίο είναι μετρήσιμο (πχ. ομάδα, η θρησκεία). Οι κατηγορίες αυτές ονομάζονται **διαβαθμίσεις**. Απαριθμώντας τις στατιστικές μονάδες με τις ίδιες διαβαθμίσεις δημιουργούμε τον **πίνακα συμπτώσεων**.

Αντίθετα τα ποσοτικά χαρακτηριστικά εκδηλώνουν ποσότητα δηλαδή είναι μεταβλητές οι οποίες παίρνουν μόνο αριθμητικές τιμές και μπορούν να ταξινομηθούν σε διακριτές (πχ. αριθμός παιδιών ανά σχολική τάξη) ή συνεχείς (πχ. το βάρος). Ένα ποσοτικό χαρακτηριστικό αν ταξινομηθεί χωρίς να χρησιμοποιηθεί κάποια κλίμακα μέτρησης αλλά χωριστεί σε διαβαθμίσεις τότε το χαρακτηριστικό μπορεί να χρησιμοποιηθεί κατά τη διάρκεια της έρευνας ως ποιοτικό χωρίς να σημαίνει ότι από τη φύση του είναι τέτοιο (Κικιλίας, et al., 2001).

Για παράδειγμα, η βαθμολογία ενός μαθητή σε ένα μάθημα είναι ποσοτικό χαρακτηριστικό. Αν όμως χρησιμοποιηθούν διαβαθμίσεις όπως φαίνεται στον παρακάτω πίνακα τότε το χαρακτηριστικό μπορεί να χρησιμοποιηθεί και ως ποιοτικό κατά τη διάρκεια της έρευνας του Στατιστικολόγου.

Πίνακας 1. Διαβαθμίσεις βαθμολογίας μαθητή

άριστη επίδοση	(όταν πάρει βαθμό Χ ≥9)
πολύ καλή επίδοση	(όταν πάρει βαθμό 6 ≤Χ<9)
καλή επίδοση	(όταν πάρει βαθμό 5≤Χ< 6)
κακή επίδοση	(όταν πάρει βαθμό Χ < 5)

Πηγή: (Champkin, 2013)

Κεφάλαιο 2: Παλινδρόμηση

2.1 Εισαγωγή στην παλινδρόμηση

Η παλινδρόμηση είναι μια διαδικασία μέσω της οποίας εξετάζεται η σχέση μεταξύ δυο ή περισσότερων μεταβλητών. Η ανάλυση παλινδρόμησης (regression analysis) εφαρμόζεται για να προβλεφθούν οι τιμές της μιας μεταβλητής (εξαρτημένης) που χρησιμοποιείται μέσω των τιμών της άλλης ή των άλλων μεταβλητών (ανεξάρτητων). Μελετώντας την ιστορική αναδρομή της Στατιστικής διαπιστώνεται ότι η ορολογία «παλινδρόμηση» χρησιμοποιήθηκε από τον Άγγλο ανθρωπολόγο Galton (1822-1911) για πρώτη φορά το 1885. Μελετήθηκε το ύψος των παιδιών σε σχέση με το ύψος των γονέων και από τα αποτελέσματα της μελέτης συμπεραίνουμε ότι τα παιδία που γεννήθηκαν από ψηλούς γονείς έχουν μια τάση κατά μέσο όρο να είναι πιο κοντά σε σχέση με τους γονείς τους, σε αντίθεση με τα παιδιά που προέρχονται από κοντούς γονείς τα οποία έχουν τη τάση να ψηλώνουν περισσότερο από τους γονείς τους (Ιωαννίδης, 2005).

Αποτελεί ενδιαφέρον στον κλάδο της Στατιστικής η εξέταση των επιδράσεων ορισμένων μεταβλητών πάνω σε ορισμένες άλλες μεταβλητές. Σκοπός είναι να εξετάζεται αν υπάρχει κάποια σχέση μέσα από την δημιουργία σχέσης εξάρτησης μεταξύ των μεταβλητών. Αυτή η σχέση που συχνά καλείται εξίσωση αποδεικνύεται απόλυτα σημαντική για να προβλεφθούν οι τιμές μιας μεταβλητής που εξετάζουμε με βάση τις γνώσεις που διαθέτουμε για τις άλλες μεταβλητές, υπό την προϋπόθεση να ισχύουν κάποιες συγκεκριμένες συνθήκες (Champkin, 2013).

Τα είδη των μεταβλητών που διακρίνουμε σε ένα πρόβλημα παλινδρόμησης είναι δυο, οι ανεξάρτητες μεταβλητές που ονομάζονται και αλλιώς ελεγχόμενες ή επεξηγηματικές και οι εξαρτημένες που ονομάζονται και μεταβλητές απόκρισης. Οι παραπάνω μεταβλητές χρησιμοποιούνται τόσο σε πειραματικές όσο και σε μη πειραματικές έρευνες. Στις πειραματικές έρευνες, η ανεξάρτητη μεταβλητή X μπορεί να ελεγχθεί, δηλαδή, έχουμε τη δυνατότητα του καθορισμού των τιμών της όπως συμβαίνει όταν θέλουμε να καθορίσουμε τη θερμοκρασία στην οποία μπορούμε να επεξεργαστούμε ένα προϊόν. Αντίθετα χρησιμοποιούμε την εξαρτημένη μεταβλητή Y με σκοπό να αντλήσουμε ένα αποτέλεσμα που σχετίζεται με τις μεταβολές των ανεξάρτητων μεταβλητών και την επίδρασή τους πάνω στην εξαρτημένη. Με άλλα λόγια εξετάζουμε την επίδραση των ανεξάρτητων μεταβλητών πάνω στην συμπεριφορά της εξαρτημένης μεταβλητής. Εξαρτημένες μεταβλητές μπορεί να είναι η επαγγελματική ικανοποίηση των εργαζομένων, το Ακαθάριστο Εγχώριο Προϊόν, η απόδοση των μαθητών σε μια τάξη κ.α (Ζαχαροπούλου, 2010).

Αντίθετα στις μη πειραματικές έρευνες (δειγματοληψίες) δεν είναι πάντοτε σαφής ο διαχωρισμός μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών διότι καμία μεταβλητή δε μπορεί να ελεγχθεί αλλά όλες είναι τυχαίες. Τέτοιες τυχαίες μεταβλητές θεωρούνται η αποδοτικότητα των μαθητών σε μια τάξη, τα χαρακτηριστικά των μαθητών, οι βαθμοί τους σε ένα διαγώνισμα, η κατάταξη ενός αριθμού προϊόντων σε ένα σουπερμάρκετ, ο αριθμός των πωλήσεων ενός προϊόντος κ.α. (Ζαχαροπούλου, 2010).

2.2 Απλή παλινδρόμηση

Με τον όρο απλή παλινδρόμηση εννοούμε την στατιστική μέθοδο που χρησιμοποιείται με σκοπό τη συσχέτιση μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Υπάρχει μια σχέση που ενώνει την εξαρτημένη μεταβλητή με τις ανεξάρτητες μεταβλητές και αυτή η σχέση είναι στατιστικής μορφής και όχι σχέση συνάρτησης. Όσον αφορά τη στατιστική σχέση που τις συνδέει μπορούμε να πούμε ότι για τις τιμές των ανεξάρτητων μεταβλητών αντιστοιχεί μια τιμή της εξαρτημένης. Αντίθετα στη σχέση εξάρτησης αντιλαμβανόμαστε ότι για κάθε τιμή της ανεξάρτητης μεταβλητής έχουμε στην εξίσωση την ίδια τιμή που αντιστοιχεί στην εξαρτημένη μεταβλητή. Δηλαδή είναι της μορφής Y=f(X), όπου Y η εξαρτημένη μεταβλητή και X η ανεξάρτητη μεταβλητή. Για λόγους διευκόλυνσης χρησιμοποιείται ο όρος «εξισώσεις παλινδρόμησης», παρά το γεγονός ότι δεν πρόκειται για εξίσωση αλλά για στατιστικό μοντέλο (Κικιλίας, et al., 2001).

Αν παραστήσουμε γραφικά τα ζεύγη (Xi, Y_i) των παρατηρήσεων 2 μεταβλητών σε ένα σύστημα ορθογώνιων αξόνων, παρατηρούμε ότι προκύπτει μια διασπορά των σημείων που αντιστοιχούν στις εξεταζόμενες μεταβλητές. Η γραφική παράσταση αυτών των σημείων ονομάζεται στικτό διάγραμμα ή διάγραμμα διασποράς (scatter diagram, scatter plot) και παρέχει σημαντικές πληροφορίες όσον αφορά τη σχέση εξάρτησης που ενδέχεται να υπάρχει μεταξύ των μεταβλητών που εξετάζουμε (Ιωαννίδης, 2005).

2.3 Πολλαπλή παλινδρόμηση

Η πολλαπλή παλινδρόμηση αποτελεί την άμεση γενίκευση της απλής παλινδρόμησης εισάγοντας περισσότερες από μια ερμηνευτικές μεταβλητές στο υπόδειγμα. Τα υποδείγματα που περιλαμβάνουν δύο ή περισσότερες ανεξάρτητες μεταβλητές ονομάζονται υποδείγματα πολλαπλής παλινδρόμησης (multiple regression models) (Χρήστου, 2007).

Σε πολλά πρακτικά καθημερινά προβλήματα χρειάζεται να χρησιμοποιήσουμε περισσότερες από δυο ανεξάρτητες μεταβλητές έτσι ώστε να προβούμε στην ερμηνεία ενός γεγονότος ή ενός φυσικού φαινομένου με μεγαλύτερη ακρίβεια. Με αυτόν τον τρόπο οδηγούμαστε σε σωστότερα και πιο ακριβή συμπεράσματα (Χρήστου, 2007).

Για παράδειγμα, για να πραγματοποιηθεί μια πρόβλεψη της ζήτησης παγωτού μιας εταιρίας σε 30 διαφορετικές μπορεί να χρησιμοποιηθεί ένα μοντέλο πολλαπλής παλινδρόμησης που να περιλαμβάνει κοινωνικοοικονομικές μεταβλητές (μέσο οικογενειακό εισόδημα, μόρφωση του αρχηγού της οικογένειας και μέσος αριθμός χρόνος εκπαίδευσης), μεταβλητές δημογραφικών χαρακτηριστικών (μέσο μέγεθος οικογενειών, ποσοστό συνταξιούχων) καθώς και περιβαλλοντολογικές εξωτερικές μεταβλητές (μέση ημερήσια θερμοκρασία, δείκτης ατμοσφαιρικής ρύπανσης). Έτσι εξάγονται ακριβή και σωστά συμπεράσματα μέσα από την ανάλυση της επίδρασης των παραπάνω μεταβλητών στο μοντέλο που εξετάζεται (Ιωαννίδης, 2005).

2.4 Γραμμική παλινδρόμηση

Η πιο απλή περίπτωση της απλής παλινδρόμησης αποτελεί η γραμμική παλινδρόμηση (simple linear regression). Στην απλή γραμμική παλινδρόμηση περιλαμβάνεται μία μόνο ανεξάρτητη μεταβλητή που συμβολίζεται με ένα *X* και η εξαρτημένη μεταβλητή που συμβολίζεται με ένα *Y*. Η εξαρτημένη μεταβλητή προσεγγίζεται ικανοποιητικά από μία γραμμική συνάρτηση του *X* δηλαδή πρόκεται για την παλινδρόμηση του παράγοντα *Y* πάνω στο παράγοντα *X* (Ιωαννίδης, 2005).

Στο Σχήμα 2.1 απεικονίζεται ένα διάγραμμα διασποράς που αποτελείται από 5 σημεία χαράσσοντας μια γραμμή, στην προκειμένη περίπτωση μια ευθεία η οποία διέρχεται από το μέσο του νέφους των σημείων αυτών. Η ευθεία αυτή είναι της εξής μορφής:

$$\hat{Y} = b_0 + b_1 X \quad (1)$$

Όπου:

- $\hat{Y}$ αποτελεί τη τιμή που έχει εκτιμηθεί ή προβλεφθεί ή έχει προσαρμοστεί δηλαδή η αναμενόμενη τιμή ή αλλιώς η εξαρτημένη τιμή για δεδομένη ανεξάρτητη μεταβλητή X.
- b_o είναι ο σταθερός όρος της εξίσωσης παλινδρόμησης
- b_1 είναι ο συντελεστής της ανεξάρτητης μεταβλητής X

Η κάθε τιμή Y_i που παίρνει η εξαρτημένη μεταβλητή προσδιορίζεται από την παρακάτω εξίσωση:

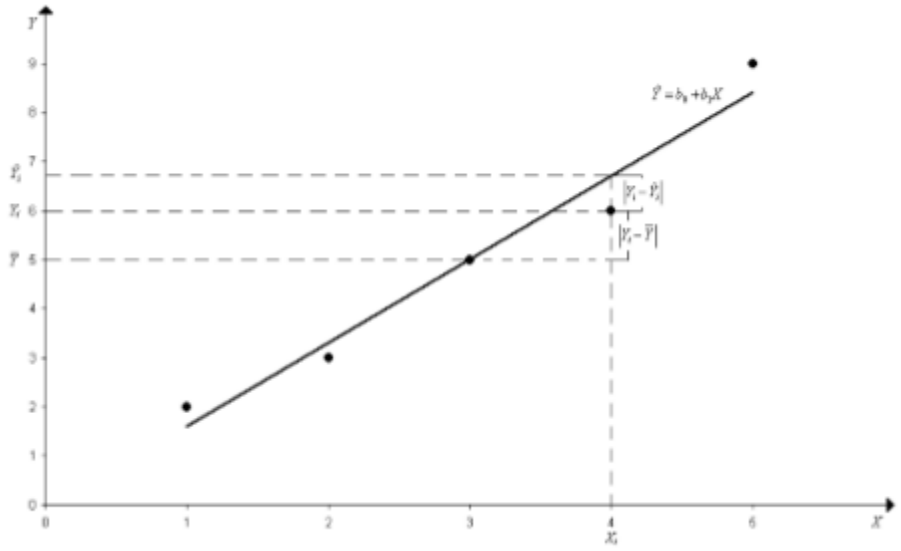
$$Y = \beta_0 + \beta_1 X + e \quad (2)$$

Όπου:

- β_0 ο σταθερός όρος και β_1 ο συντελεστής παλινδρόμησης της ανεξάρτητης μεταβλητής X
- e το τυπικό σφάλμα ή κατάλοιπο που υπολογίζεται από τη διαφορά $|Y - \hat{Y}|$.

Τα b_0 και b_1 στην ευθεία παλινδρόμησης αποτελούν τους εκτιμητές των συντελεστών παλινδρόμησης (regression estimators) β_0 και β_1 (Κικιλίας, et al., 2001).

Σχήμα 2.1: Ευθεία απλής γραμμικής παλινδρόμησης

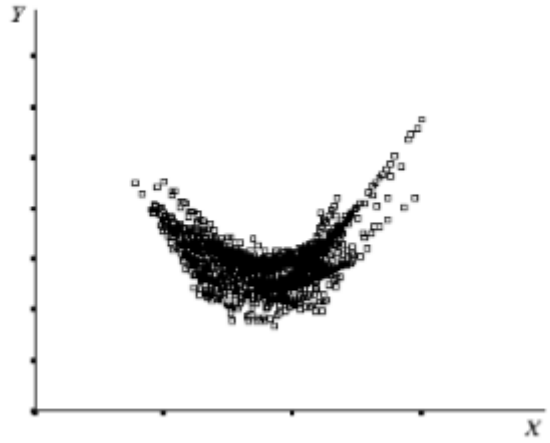


Πηγή: (Ιωαννίδης, 2005)

2.5 Μη γραμμική παλινδρόμηση

Στην περίπτωση που η γραμμή παλινδρόμησης που περνάει από το μέσο του νέφους των τιμών ενός διαγράμματος διασποράς δεν είναι ευθεία, τότε είναι αναγκαίο να εκτιμηθεί μια γραμμή μη γραμμικής παλινδρόμησης (nonlinear regression). Η μη γραμμική παλινδρόμηση απεικονίζεται με το διάγραμμα διασποράς στο Σχήμα 2.2 (Χρήστου, 2007).

Σχήμα 2.2: Διάγραμμα διασποράς μη γραμμικής παλινδρόμησης

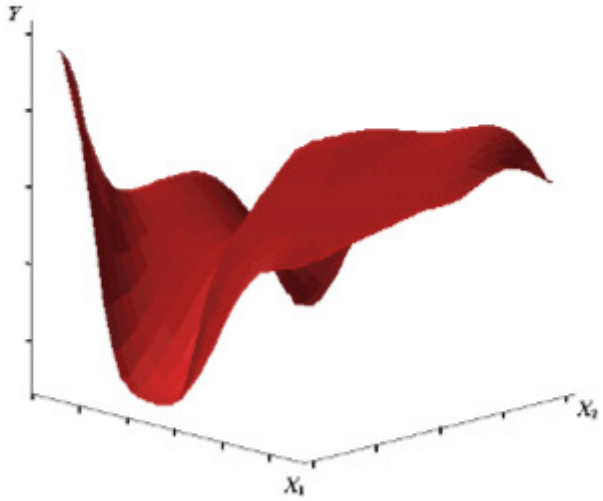


Πηγή: (Ιωαννίδης, 2005)

Όταν οι ανεξάρτητες μεταβλητές είναι περισσότερες από μια, δηλαδή στην περίπτωση της πολλαπλής παλινδρόμησης το διάγραμμα διασποράς είναι ν-διάστατης μορφής που περιλαμβάνει ν διαστάσεις οι οποίες είναι ίσες με τις ανεξάρτητες μεταβλητές συν άλλη μια. Το διάγραμμα διασποράς ονομάζεται επιφάνεια

παλινδρόμησης ή διαφορετικά επιφάνεια απόκρισης. Στο Σχήμα 2.3 παρουσιάζεται ένα τρισδιάστατο διάγραμμα που περιλαμβάνει 2 ανεξάρτητες μεταβλητές X_1 και X_2 (Ζαχαροπούλου, 2010).

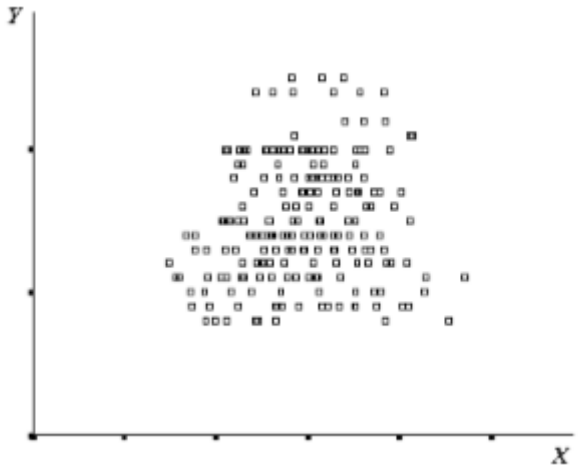
Σχήμα 2.3: Τρισδιάστατο διάγραμμα διασποράς



Πηγή: (Χρήστου, 2007)

Τέλος, στο Σχήμα 2.4 απεικονίζεται ένα διάγραμμα διασποράς, όπου δεν εντοπίζεται καμία συσχέτιση μεταξύ των 2 μεταβλητών. Αυτό το γεγονός μας οδηγεί στο συμπέρασμα ότι πρόκειται για μια μη γραμμική παλινδρόμηση (Χρήστου, 2007).

Σχήμα 2.4: Μη ύπαρξη σχέσης μεταξύ 2 μεταβλητών



Πηγή: (Χρήστου, 2007)

Κεφάλαιο 3: Απλή γραμμική παλινδρόμηση

3.1 Βασικές έννοιες

Όπως έχουμε προαναφέρει σκοπός της απλής γραμμικής παλινδρόμησης είναι να προβλέψουμε τη συναρτησιακή σχέση μεταξύ δυο μεταβλητών. Ας θεωρήσουμε δύο μεταβλητές X ,Y. Στην περίπτωση που αυτές οι δυο μεταβλητές συνδέονται με μια σχέση η οποία έχει την μορφή:

$$\boldsymbol{Y = f(X)}\ \textbf{(3)}$$

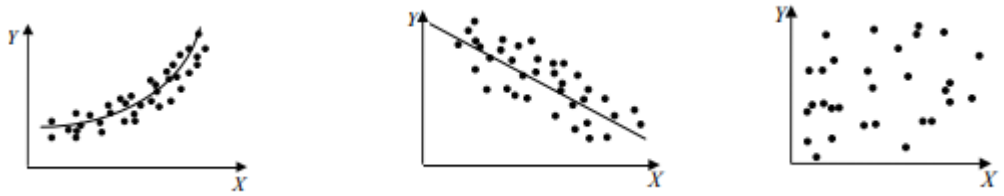
τότε μπορούμε μέσα από αυτή τη μορφή για κάθε τιμή της X να προβλέψουμε την ακριβή τιμή της Y. Αυτό σημαίνει ότι αν υποθέσουμε ότι οι τιμές της Y δεν υπόκεινται σε σφάλματα, τότε οδηγούμαστε στο συμπέρασμα ότι οι δύο μεταβλητές συνδέονται μεταξύ τους με τη παραπάνω **συναρτησιακή-προσδιοριστική (deterministic) σχέση $\boldsymbol{Y = f(X)}$** (Ζαχαροπούλου, 2010)**.**

Για να κατανοήσουμε καλύτερα αυτή τη συναρτησιακή σχέση μπορούμε να εξετάσουμε την περίπτωση που κάποιος καταθέτει ένα ποσό στο Ταμιευτήριο και το τόκο που παίρνει για το ποσό αυτό. Αυτές οι δυο μεταβλητές συνδέονται μεταξύ τους με μια συναρτησιακή-προσδιοριστική σχέση. Σε αυτή την περίπτωση διαπιστώνεται ότι όλα τα σημεία που βρίσκονται πάνω στη καμπύλη στο διάγραμμα διασποράς είναι της μορφής $Y = f(X)$. Επαναλαμβάνοντας το πείραμα όσες φορές θέλουμε, θα διαπιστώσουμε ότι όταν θέτουμε $X = x_i$, το αποτέλεσμα που θα παίρνουμε πάντα είναι ίδια ακριβώς τιμή για την εξαρτημένη μεταβλητή Y (Χρήστου, 2007).

Οι σχέσεις μεταξύ μεταβλητών που δεν προσδιορίζονται λέγονται στοχαστικές-στατιστικές σχέσεις. Επαναλαμβάνοντας αρκετές φορές το πείραμα και αν θέσουμε $X = x_i$ τότε θα δούμε ότι η τιμή x_i της X δεν αντιστοιχίζεται με μόνο μια τιμή y_i της Y. Αντίθετα της αναλογεί ένας μεγάλος αριθμός από διαφορετικές τιμές της εξαρτημένης μεταβλητής Y. Ας πάρουμε το παράδειγμα της ζήτησης προϊόντων ώστε να κατανοηθεί καλύτερα. Έστω ότι η τιμή X αποτελεί τη τιμή ενός παγωτού και Y είναι η ζήτηση του παγωτού, η Y εξαρτάται και επηρεάζεται από την ανεξάρτητη μεταβλητή X δηλαδή η ζήτηση του παγωτού εξαρτάται από τη τιμή του, υπάρχει μια στοχαστική σχέση μεταξύ τους. Αυτό συμβαίνει διότι η ζήτηση ενός προϊόντος επηρεάζεται από ένα μεγάλο εύρος παραγόντων όπως είναι το εισόδημα του καταναλωτή, οι τιμές των ομοειδών προϊόντων, οι προτιμήσεις του καταναλωτή και οι καταναλωτικές του συνήθειες κλπ (Ιωαννίδης, 2005).

Χρησιμοποιούμε το **διάγραμμα διασποράς** με σκοπό να απεικονίσουμε την σχέση μεταξύ κάποιων μεταβλητών. Γενικά σε μια στοχαστική σχέση το διάγραμμα διασποράς αποτελεί ένα νέφος σημείων το οποίο σε πολλές περιπτώσεις προσδιορίζει μια ιδεατή γραμμή η οποία προσδίδει μια πρώτη εικόνα για τη σχέση που συνδέει τις δυο εξεταζόμενες μεταβλητές. Όταν τα σημεία διαγράμματος διασποράς βρίσκονται πάρα πολύ κοντά στην ιδεατή γραμμή τότε η σχέση μεταξύ των δυο μεταβλητών είναι όλο και πιο ισχυρή. Παρακάτω παρουσιάζονται τρία διαγράμματα με σκοπό την καλύτερη κατανόηση του διαγράμματος διασποράς, όπου εξετάζονται τρεις περιπτώσεις: α) ισχυρή σχέση, β) λιγότερο ισχυρή σχέση και γ) καμία σχέση μεταξύ των δυο μεταβλητών (Χρήστου, 2007).

Σχήμα 3.1: Διαγράμματα διασποράς



Πηγή: (Ιωαννίδης, 2005)

Στο πρώτο διάγραμμα διασποράς παρατηρούμε ότι τα σημεία βρίσκονται πολύ κοντά στην ιδεατή γραμμή γεγονός που αποδεικνύει την ισχυρή σχέση στην οποία με την αύξηση των τιμών της *X* αυξάνονται γενικά και οι τιμές της *Y*. Στο δεύτερο διάγραμμα διασποράς παρατηρούμε μια λιγότερο ισχυρή σχέση όπως βλέπουμε τα σημεία είναι πιο απομακρυσμένα από την ιδεατή γραμμή. Αυτό αποδεικνύει ότι με την αύξηση των τιμών της *X* οδηγούμαστε σε μείωση των τιμών της *Y*. Στο τρίτο διάγραμμα διασποράς δεν φαίνεται να υπάρχει κάποια συναρτησιακή σχέση μεταξύ των δυο μεταβλητών Χ και Υ (Κικιλίας, et al., 2001). Τα σημεία όπως απεικονίζονται στο διάγραμμα είναι διάσπαρτα στο χώρο αποδεικνύοντας ότι δεν υπάρχει σχέση αιτιότητας μεταξύ των μεταβλητών που εξετάζουμε.

Κατά κύριο λόγο όπως προαναφέραμε σε προηγούμενη ενότητα, δύο μεταβλητές που αλληλεξαρτώνται δηλαδή συνδέονται είτε με μια συναρτησιακή-προσδιοριστική σχέση είτε με μια στοχαστική σχέση ονομάζονται **«εξαρτημένες»**. Στην περίπτωση που υπάρχει εξάρτηση μεταξύ δύο μεταβλητών, τότε μπορούμε να προσδιορίσουμε τη μια μεταβλητή από αυτές ως «αιτία» και την άλλη μεταβλητή ως «αποτέλεσμα» (Montgomery, et al., 2012). Δηλαδή υπάρχει μια σχέση αιτίας και αιτιατού. Αυτό όμως συμβαίνει στην περίπτωση μόνο που η εξάρτηση αυτή οφείλεται σε σχέση αιτιότητας μεταξύ των δύο μεταβλητών και όχι σε μια απλή συμμεταβολή η οποία μπορεί να οφείλεται στην εξάρτηση των δύο μεταβλητών που προέρχεται από μια τρίτη μεταβλητή. Για παράδειγμα στην περίπτωση που η μεταβλητή Χ είναι το ετήσιο εισόδημα μιας οικογένειας και οι μεταβλητές Υ, Ζ είναι τα ποσά που ξοδεύει σε ετήσια βάση η οικογένεια αυτή για ψωμί και για αγορά εφημερίδων τότε υπάρχουν δυο ενδεχόμενα. Το πρώτο ενδεχόμενο είναι ότι στην περίπτωση διαπίστωσης ύπαρξης σχέσης ανάμεσα στα Χ και Υ σε έναν αριθμό οικογενειών τότε καταλήγουμε στο συμπέρασμα ότι οι 2 μεταβλητές εξαρτώνται η μια από την άλλη και έτσι μπορούμε να ονομάσουμε την ανεξάρτητη μεταβλητή Χ ως «αιτία» και την εξαρτημένη μεταβλητή Υ ως «αποτέλεσμα». Το δεύτερο ενδεχόμενο περιλαμβάνει την περίπτωση που αν διαπιστωθεί σχέση μεταξύ των Υ και Ζ, γεγονός που είναι πολύ πιθανό, εφόσον και οι δύο μεταβλητές μεταβάλλονται καθώς μεταβάλλεται το ετήσιο εισόδημα Χ, οδηγούμαστε στο εισόδημα ότι πρόκειται για «νόθα» εξάρτηση (Χρήστου, 2007).

Προκειμένου να μελετήσουμε τη στοχαστική εξάρτηση και σχέση μεταξύ των μεταβλητών Χ και Υ επιδιώκουμε να διερευνήσουμε μια σχέση η οποία θα συνδέει τις 2 μεταβλητές αλλά σε αυτή τη περίπτωση δεν θα πάρουμε την ακριβή σχέση αλλά την προσεγγιστική. Θα έχουμε δηλαδή μια προσεγγιστική εικόνα της εξάρτησης μεταξύ των μεταβλητών Χ και Υ που απεικονίζεται στο διάγραμμα διασποράς ως

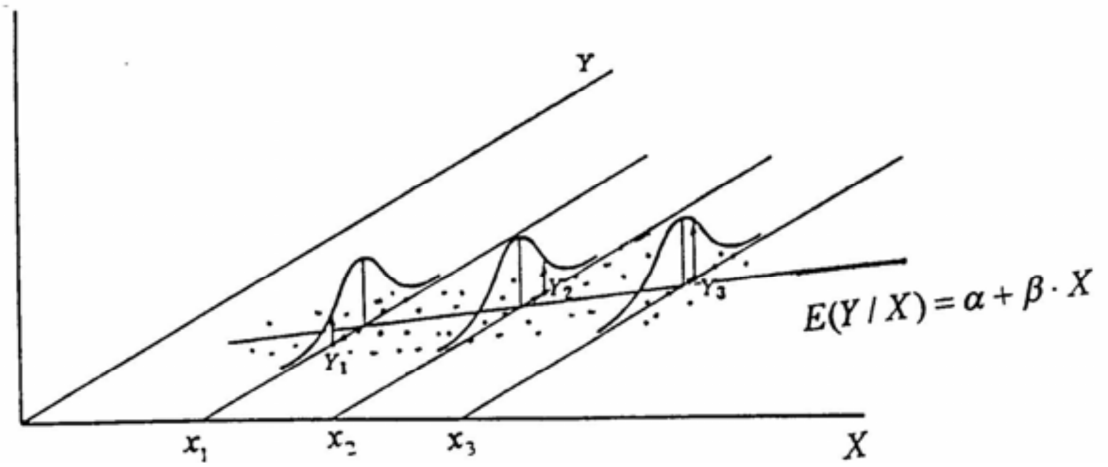
σημεία που δε βρίσκονται πάνω αλλά γύρω από μια καμπύλη (Montgomery, et al., 2012). Η μέθοδος που χρησιμοποιείται στην ανάλυση παλινδρόμησης για να περιγράψει την στοχαστική εξάρτηση μεταξύ δυο μεταβλητών ονομάζεται **μέθοδος των ελαχίστων τετραγώνων** η οποία θα αποσαφηνιστεί διεξοδικά σε επόμενη ενότητα και θα χρησιμοποιηθεί για να μελετήσουμε την γραμμική μορφή παλινδρόμησης η οποία αποτελεί την πιο απλή μορφή στοχαστικής εξάρτησης (Παπαδόπουλος, 2008).

3.2 Διάγραμμα διασποράς

Το διάγραμμα διασποράς αποτελεί τον πιο απλό τρόπο προκειμένου να διαπιστωθεί αν υπάρχει ή όχι συσχέτιση μεταξύ δυο μεταβλητών. Στη περίπτωση που το διάγραμμα διασποράς μεταξύ δυο μεταβλητών Χ και Υ είναι της μορφής J τότε λέμε ότι η σχέση μεταξύ των μεταβλητών είναι προσεγγιστικά γραμμική. Στη γραμμική σχέση που συνδέει δυο μεταβλητές λέμε ότι εξετάζουμε μια πολύ απλή μορφή παλινδρόμησης στην οποία έχουμε μόνο δυο μεταβλητές, την ανεξάρτητη μεταβλητή Χ και την εξαρτημένη μεταβλητή Υ. Η εξαρτημένη μεταβλητή υπάρχει η δυνατότητα να προσεγγισθεί μέσα από τη γραμμική συνάρτηση της ανεξάρτητης μεταβλητής (Παπαδόπουλος, 2008).

Η σχέση $Y = \alpha + \beta X$ είναι γραμμικής μορφής και είναι δυνατό να δώσει πλήρη περιγραφή για τη γραμμική στοχαστική εξάρτηση που συνδέει τις δύο μεταβλητές X και Y. Προκειμένου να το κατανοήσουμε αυτό βλέπουμε ότι στο παράδειγμα της ζήτησης, η μεταβλητή X είναι η τιμή ενός προϊόντος ενώ η τιμή Y είναι η ζήτηση του προϊόντος αυτού, αν διατηρήσουμε τη τιμή X στο ίδιο επίπεδο $X = x_1$ διαπιστώνουμε ότι οι τιμές που αντιστοιχούν στην Y θα διαφέρουν πραγματοποιώντας διαφορετικές επαναλήψεις. Ενώ σε μια άλλη περίπτωση όπου το X είναι η ποσότητα λιπάσματος και Y είναι η απόδοση μιας καλλιέργειας, διατηρώντας πάλι τη τιμή X στο ίδιο επίπεδο $X = x_1$ τότε διαπιστώνουμε ότι οι αντίστοιχες τιμές του Y θα είναι φυσικά διαφορετικές στις διάφορες επαναλήψεις εφόσον επηρεάζουν το έδαφος άλλοι εξωγενείς παράγοντες όπως είναι η θερμοκρασία, οι βροχοπτώσεις, η ποιότητα του εδάφους κ.α. Επιπλέον, σημειώνονται τα λεγόμενα «σφάλματα μέτρησης των τιμών της Y» που οφείλονται στην ελλιπή πληροφόρηση. Κατά συνέπεια για $X = x_1$ το Y που του αντιστοιχεί αποτελεί μια «τυχαία μεταβλητή Y_1» η οποία ακολουθεί κάποια κατανομή. Αντίστοιχη περίπτωση είναι η $X = x_2$ όπου διαπιστώνεται άλλη κατανομή για Y_2 κ.ό.κ (Ιωαννίδης, 2005). Τα παραπάνω απεικονίζονται στο παρακάτω διάγραμμα απλής γραμμικής παλινδρόμησης.

Σχήμα 3.2: Γραμμικό διάγραμμα διασποράς



Πηγή: (Χρήστου, 2007)

Συνεπώς στην εξίσωση Y =α + β · X , πρέπει να προστεθεί άλλος ένας όρος που συμβολίζεται με ε όπου για δεδομένο X, συμβολίζει τη διαφορά που προκύπτει μεταξύ της παρατηρούμενης και της θεωρητικής τιμής της μεταβλητής Y (α + β · X) τιμή της Y. Δηλαδή, ε = Y − (α + β · X). Με αυτό τον τρόπο έχουμε το στοχαστικό μοντέλο το οποίο είναι της μορφής:

$$\boldsymbol{Y = \alpha + \beta X + \varepsilon} \quad \textbf{(4)}$$

Για καθαρά απλουστευτικούς λόγους των υπολογισμών και της επίλυσης του προβλήματος, κάνουμε κάποιες υποθέσεις, όπως E(ε) = 0 και E(Y / X) =α + β · X. Με άλλα λόγια κάνουμε την υπόθεση ότι η μέση τιμή των σφαλμάτων είναι μηδέν και ότι για διάφορες τιμές της X, οι μέσες τιμές Y που αντιστοιχούν σε αυτές, βρίσκονται πάνω σε μια ευθεία. Η εν λόγω ευθεία E(Y / X) =α + β · X ονομάζεται πληθυσμιακή ευθεία παλινδρόμησης (Champkin, 2013).

3.3 Μέθοδος ελαχίστων τετραγώνων

Η μέθοδος των ελαχίστων τετραγώνων χρησιμοποιείται για να εκτιμηθεί η σχέση $\hat{Y} = \hat{\alpha} + \hat{\beta}$ X της ευθείας E(Y / X) =α + β · X (όπου $\hat{\alpha}$ και $\hat{\beta}$ εκτιμήτριες των α και β αντίστοιχα). Η εκτίμηση της παραπάνω σχέσης ονομάζεται ευθεία ελαχίστων τετραγώνων λόγω της μεθόδου υπολογισμού των παραμέτρων της. Στόχος της μεθόδου είναι να προσδιορίσει τους σταθερούς συντελεστές (Χρήστου, 2007).

Η μέθοδος των ελαχίστων τετραγώνων χρησιμοποιείται για να περιγράψει ένα φαινόμενο όταν είναι γνωστές οι πειραματικές τιμές των μεγεθών που μελετώνται αλλά όχι η ακριβή σχέση που υπάρχει μεταξύ τους. Δηλαδή γίνεται μια προσπάθεια προσδιορισμού της άγνωστης σχέσης μεταξύ των μεγεθών η οποία ταιριάζει καλύτερα στα πειραματικά δεδομένα που εξετάζονται βασιζόμενοι σε μια σειρά από γνωστές σχέσεις. Αυτό το σκοπό καλείται να καλύψει η εν λόγω μέθοδος (Montgomery, et al., 2012).

Συγκεκριμένα με βάση τα παραπάνω η μέθοδος οδηγεί στις εκτιμήσεις των παραμέτρων α και β που ελαχιστοποιούν το άθροισμα των τετραγωνικών αποκλίσεων των παρατηρούμενων τιμών y_i της Y από τις τιμές $\alpha + \beta \cdot X_i$ που εκτιμάται ότι θα έπρεπε να παρατηρούνται για την Y (με βάση την υπόθεση που έχει προηγηθεί $E(Y|X=x) = \alpha + \beta \cdot X$). Έτσι αν θεωρήσουμε ν ζεύγη παρατηρήσεων (x_i, y_i) = 1,2,3,...,ν η προσέγγιση που εξετάζουμε έχει την εξής μορφή:

$$\boldsymbol{y_i = \alpha + \beta \cdot x_i + \varepsilon_i} \quad \textbf{(5)}$$

όπου:

- ε_i παριστάνουν τις αποκλίσεις της πραγματικής τιμής y_i από την προσαρμοσμένη (θεωρητική) $\alpha + \beta \cdot x_i$. Δηλαδή, $\varepsilon_i = y_i - (\alpha + \beta\, x_i)$.

Όπως αντιλαμβανόμαστε είναι απαραίτητη η εκτίμηση των παραμέτρων α και β έτσι ώστε να ελαχιστοποιηθούν οι ποσότητες ε_i. Έτσι αναζητούνται οι τιμές των παραμέτρων για τις οποίες οδηγούμαστε στην ελαχιστοποίηση των τετραγώνων των ε_i. Η ποσότητα που αναζητείται είναι η εξής:

$$\sum_{i=1}^{\nu} \varepsilon_i^2 = \sum_{i=1}^{\nu} (y_i - \alpha - \beta\, x_i)^2 \quad \textbf{(6)}$$

Στην παραπάνω σχέση συναντάμε έναν περιορισμό ως προς την ασφάλεια του κριτηρίου επιλογής του συγκεκριμένου αθροίσματος αφού κάποια αρνητικά ε_i μπορεί να οδηγήσουν σε θετικές ποσότητες του αθροίσματος (Χρήστου, 2007). Στη συνέχεια παραγωγίζουμε την εξίσωση (6) ως προς τις παραμέτρους α και β και θέτοντας ίσο με το μηδέν οδηγούμαστε στις ακόλουθες κανονικές εξισώσεις:

$$\sum_{i=1}^{\nu} y_i = \nu \cdot a + \beta \cdot \sum_{i=1}^{\nu} x_i \quad \textbf{(7)}$$

$$\sum_{i=1}^{\nu} x_i\, y_i = a \cdot \sum_{i=1}^{\nu} x_i + \beta \sum_{i=1}^{\nu} \cdot\, x_i^2 \quad \textbf{(8)}$$

Λύνοντας το σύστημα των παραπάνω εξισώσεων έχουμε την εξής εξίσωση:

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} \text{ και } \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \quad \textbf{(9)}$$

Συνεπώς η εκτίμηση των ελαχίστων τετραγώνων $\hat{Y} = \hat{\alpha} + \hat{\beta}\, X$ της ευθείας παλινδρόμησης από το δείγμα των ν ζευγών παρατηρήσεων είναι η εξής:

$$\hat{Y} = \hat{a} + \hat{\beta} \cdot X = \bar{y} - \hat{\beta} \cdot \bar{x} + \hat{\beta} \cdot X = \bar{y} + \hat{\beta}\,(X - \bar{x})$$

Ή

$$\hat{Y} = \bar{y} + \frac{S_{xy}}{S_x^2} \cdot (X - \bar{x}) \quad \textbf{(10)}$$

Από τη σχέση (10) βλέπουμε ότι η *ευθεία ελαχίστων τετραγώνων*, διέρχεται από το σημείο $(\bar{x}, \bar{y})$. Είναι σημαντικό να σημειωθεί ότι θα πρέπει να πραγματοποιείται διάκριση μεταξύ της παρατηρούμενης τιμής του Y και της $\hat{Y}$ που εκτιμάται. Η

παρατηρούμενη τιμή $\mathbf{y_i}$ αποτελεί τη πραγματική τιμή της *Y* ενώ η τιμή $\hat{y}_i$ της $\hat{Y}$ αποτελεί εκτίμηση της μέσης τιμής $E(Y|X=x_i)$ (Montgomery, et al., 2012).

3.4 Συντελεστής γραμμικής συσχέτισης του Pearson

Ο συντελεστής γραμμικής συσχέτισης του Pearson συμβολίζεται με το σύμβολο r και αποτελεί ένα μέτρο μεγέθους της εξέτασης της γραμμικής συσχέτισης μεταξύ δύο μεταβλητών. Ο τύπος από τον οποίο ορίζεται είναι ο εξής:

$$\mathbf{r}=\frac{\boldsymbol{Sxy}}{\mathbf{Sx}*\mathbf{Sy}} \qquad \mathbf{(11)}$$

Η παραπάνω εξίσωση είναι αποτέλεσμα των παρακάτω πράξεων:

$$\boldsymbol{Sxy} = \mathbf{Cov}\,(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{i=1}^{v}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})\cdot\,(\boldsymbol{y_i}-\bar{\boldsymbol{y}})}{\boldsymbol{v}=\mathbf{1}} = \frac{\sum_{i=1}^{v}(\boldsymbol{x_i y_i})-v\cdot\,\bar{x}\,\cdot\bar{y}}{\boldsymbol{v}=\mathbf{1}}$$

$$\boldsymbol{Sx} = \sqrt{\frac{\mathbf{1}}{v-1}\sum_{i=1}^{v}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^{\mathbf{2}}}\ \textbf{και}\ \boldsymbol{Sy} = \sqrt{\frac{\mathbf{1}}{v-1}\sum_{i=1}^{v}(\boldsymbol{y_i}-\bar{\boldsymbol{y}})^{\mathbf{2}}} \quad \mathbf{(12)}$$

Έτσι καταλήγουμε στην εξής εξίσωση:

$$\mathbf{r}=\frac{\boldsymbol{Sxy}}{\mathbf{Sx}*\mathbf{Sy}} = \frac{\sum_{i=1}^{v}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})\cdot(\boldsymbol{y_i}-\bar{\boldsymbol{y}})}{\sqrt{\sum_{i=1}^{v}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^{\mathbf{2}}}\cdot\sqrt{\sum_{i=1}^{v}(\boldsymbol{y_i}-\bar{\boldsymbol{y}})^{\mathbf{2}}}} = \frac{\sum_{i=1}^{v}\boldsymbol{x_i y_i}-v\cdot\bar{\chi}\bar{y}}{\sqrt{\sum_{i=1}^{v}\boldsymbol{x_i^2}-v\bar{x}^2}\sqrt{\sum_{i=1}^{v}\boldsymbol{y_i^2}-v\bar{y}^2}}$$

(13)

Η εξίσωση **(13)** παρουσιάζει τον πληθυσμιακό συντελεστή γραμμικής συσχέτισης του Pearson και παίρνει τιμές που περιλαμβάνονται στο διάστημα [-1, 1]. Στην περίπτωση που r = ±1 υπάρχει τέλεια γραμμική συσχέτιση μεταξύ των μεταβλητών που εξετάζουμε. Στην περίπτωση που $-0{,}3 \le r < 0{,}3$ δεν υπάρχει γραμμική συσχέτιση. Το γεγονός αυτό όμως δεν σημαίνει ότι δεν υπάρχει άλλου είδους συσχέτιση μεταξύ των δυο μεταβλητών που εξετάζουμε. Επιπλέον ισχύουν οι παρακάτω περιπτώσεις:

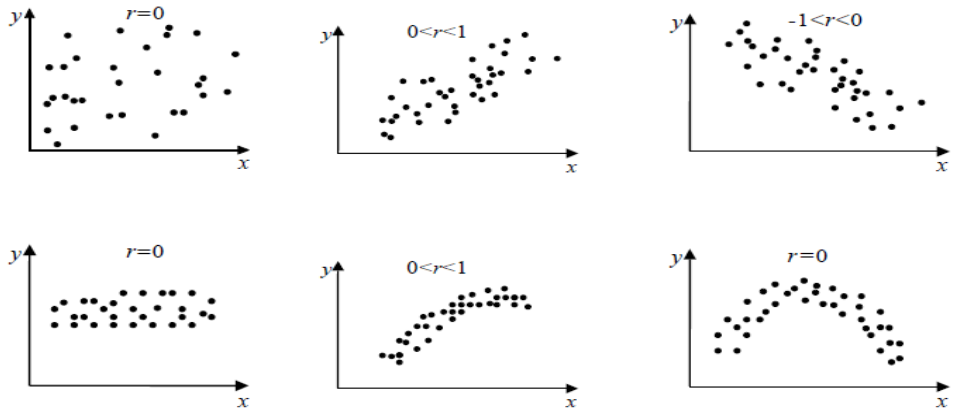
- Αν $-0{,}5 < r \le -0{,}3$ ή $0{,}3 \le r < 0{,}5$ υπάρχει ασθενής γραμμική συσχέτιση.
- Αν $-0{,}7 < r \le -0{,}5$ ή $0{,}5 \le r < 0{,}7$ υπάρχει μέση γραμμική συσχέτιση.
- Αν $-0{,}8 < r \le -0{,}7$ ή $0{,}7 \le r < 0{,}8$ υπάρχει ισχυρή γραμμική συσχέτιση.
- Αν $-1 < r \le -0{,}8$ ή $0{,}8 \le r < 1$ υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

Στην περίπτωση που παρουσιάζονται θετικές τιμές του *r* δεν συνεπάγεται απαραίτητα μεγαλύτερο βαθμό γραμμικής συσχέτισης σε σχέση με το βαθμό γραμμικής συσχέτισης που εμφανίζουν αρνητικές τιμές του συντελεστή Pearson. Δηλαδή, ο βαθμός γραμμικής συσχέτισης δεν καθορίζεται από το πρόσημο του συντελεστή αλλά από την απόλυτη τιμή του r. Μας παρέχει πληροφορίες για το αν η αύξηση της μιας μεταβλητής οδηγεί σε αύξηση ή μείωση της άλλης. Για παράδειγμα αν η τιμή του συντελεστή είναι -0,8 τότε εμφανίζει πιο ισχυρή γραμμική συσχέτιση συγκριτικά με την τιμή του συντελεστή 0,7. Για τις τιμές r=-0,7 και r=0.7

συμπεραίνουμε ότι παρουσιάζουν τον ίδιο βαθμό γραμμικής συσχέτισης αλλά αντίθετο είδος (Ιωαννίδης, 2005).

Για να υπολογίσουμε το συντελεστή γραμμικής συσχέτισης r θα πρέπει να λάβουμε υπόψιν μας μόνο τις περιπτώσεις που το διάγραμμα διασποράς έχει σχήμα επιμήκους κεκλιμένης έλλειψης ή πλατυσμένου J. Στην περίπτωση όμως που το διάγραμμα διασποράς έχει άλλη μορφή τότε η τιμή του συντελεστή θα είναι πολύ μικρή. Θα υπάρχει συσχέτιση αλλά δεν θα είναι γραμμική. Δηλαδή είναι πολύ πιθανό να υπάρχει μεγάλη μη γραμμική συσχέτιση.

Σχήμα 3.3. Διαγράμματα διασποράς συντελεστή γραμμικής συσχέτισης r



Πηγή: (Παπαδόπουλος, 2008)

Ο συντελεστής r δεν χρησιμοποιείται σε πειραματικές έρευνες όπου οι τιμές της μίας μεταβλητής ελέγχονται από τον ερευνητή. Χρησιμοποιείται όπως προαναφέρθηκε σαν ένας εκτιμητής του πληθυσμιακού συντελεστή γραμμικής συσχέτισης ρ μόνο στην περίπτωση που τα ζεύγη $(x_1, y_1),\ldots,(x_v, y_v)$ προέρχονται από τυχαία δειγματοληψία. Επίσης δεν πρέπει να συγχέουμε την έννοια της συσχέτισης με αυτή της αιτιότητας. Όταν έχουμε μια μη πειραματική έρευνα όπως είναι η δειγματοληψία και εξετάζουμε δυο μεταβλητές οι οποίες φαίνεται ότι είναι συσχετισμένες τότε λέμε ότι αυτές οι δυο μεταβλητές συνδέονται μεταξύ τους με κάποια σχέση. Συνεπώς δεν υπάρχει πάντα αιτιότητα. Ίσως συνδέονται με σχέση αιτιότητας ίσως όχι. Υπάρχει επίσης το ενδεχόμενο να επηρεάζονται από μια άλλη μεταβλητή τρίτη (Montgomery, et al., 2012).

Για παράδειγμα ας υποθέσουμε ότι το βάρος των φοιτητών μιας σχολής, ηλικίας 18 έως 25 ετών παρουσιάζει ισχυρή θετική γραμμική συσχέτιση με την αντιληπτική ικανότητα των φοιτητών. Όπως είναι λογικό το ύψος δεν σχετίζεται με την αντιληπτική ικανότητα των φοιτητών και το αντίθετο. Αλλά τόσο το ύψος όσο και η πνευματική αντίληψη των φοιτητών επηρεάζονται από άλλους, εξωτερικούς παράγοντες (Παπαδόπουλος, 2008).

Συνεπώς γίνεται κατανοητό ότι απαιτείται προσοχή όσον αφορά την ερμηνεία και τη χρήση του συντελεστή συσχέτισης r διότι πολλές φορές μπορεί να μας οδηγήσει σε παρερμηνείες και λανθασμένα συμπεράσματα. Πάντα για την εξαγωγή αιτιολογικών συμπερασμάτων απαιτείται πειραματισμός. Συνεπώς κατά την εξέταση αιτιώδους σχέσης μεταξύ δυο μεταβλητών δεχόμαστε ότι υπάρχει αλληλεξάρτηση μόνον όταν βασίζεται σε λογικά και επιστημονικά κριτήρια.

Τέλος, αξίζει να αναφερθεί και η συνδιασπορά δυο μεταβλητών X και Y που συμβολίζεται με S_{xy} και ονομάζεται δειγματική συνδιασπορά. Αυτό το μέτρο δεν χρησιμοποιείται για να συσχετίσει δυο μεταβλητές γιατί επηρεάζεται από τις μονάδες στις οποίες εκφράζονται αυτές οι μεταβλητές. Η πληθυσμιακή συνδιασπορά ορίζεται ανάλογα με τη S_{xy} αλλά συμβολίζεται με σ_{xy}. Είναι υπεύθυνη για τη συμμεταβολή-συσχέτιση δύο μεταβλητών μέσω του αθροίσματος των γινομένων των αποκλίσεων των τιμών τους από τους αντίστοιχους μέσους. Όταν παίρνει μεγάλες τιμές σημαίνει ότι υπάρχει συμμεταβολή-συσχέτιση ενώ οι μικρές τιμές της υποδηλώνουν ότι δεν παρουσιάζεται συμμεταβολή-συσχέτιση (Κικιλίας, et al., 2001).

3.5 Συντελεστής προσδιορισμού

Στην απλή γραμμική παλινδρόμηση ο συντελεστής προσδιορισμού μεταξύ 2 μεταβλητών *X, Y* δείχνει το ποσοστό της μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής *Y* που εξηγείται από την ανεξάρτητη μεταβλητή *X*. Ο συντελεστής προσδιορισμού συμβολίζεται με το R^2 (Champkin, 2013).

Για παράδειγμα στην περίπτωση που εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων σε μια απλή γραμμική παλινδρόμηση και έχουμε σαν αποτέλεσμα R^2= 60% τότε μπορούμε να διαπιστώσουμε ότι το 60% της μεταβλητότητας των τιμών της εξαρτημένης μας μεταβλητής εξηγείται από τις τιμές της ανεξάρτητης μεταβλητής ενώ το υπόλοιπο 40% του δείγματος που εξετάζουμε δεν ερμηνεύεται από τις τιμές της ανεξάρτητης μεταβλητής αλλά μπορεί να οφείλεται σε τυχαία σφάλματα. Δηλαδή, η εξαρτημένη μεταβλητή *Y* μεταβάλλεται κατά 60% λόγω της μεταβολής της ανεξάρτητης μεταβλητής *X* (Χρήστου, 2007).

Ας δούμε στο σημείο αυτό από πού προκύπτει ο συντελεστής προσδιορισμού. Από τη σχέση $y_i-\bar{y}=(y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{y})$ μπορούμε να δείξουμε:

$$\sum_{i=1}^{\nu}(y_i-\bar{y})^2=\sum_{i=1}^{\nu}(y_i-\hat{Y}_i)^2+\sum_{i=1}^{\nu}(\hat{Y}_i-\bar{y})^2 \quad \textbf{(13)}$$

Το άθροισμα που προκύπτει από τη παραπάνω σχέση είναι:

$$\textbf{SSTO}=\sum_{i=1}^{\nu}(y_i-\bar{y})^2 \qquad \textbf{(14)}$$

Ονομάζεται ολικό άθροισμα τετραγώνων ή αλλιώς ολική μεταβλητότητα των yi και όπως φαίνεται από τη σχέση (13) μπορούμε να την αναλύσουμε σε δύο συνιστώσες οι οποίες είναι το άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares) και το άθροισμα τετραγώνων των σφαλμάτων (error sum of squares). Τα παραπάνω αθροίσματα φαίνονται από τις παρακάτω εξισώσεις αντίστοιχα (Montgomery, et al., 2012).

$$\textbf{SSR}=\sum_{i=1}^{\nu}(\hat{Y}_i-\bar{y})^2 \textbf{ (15)}$$

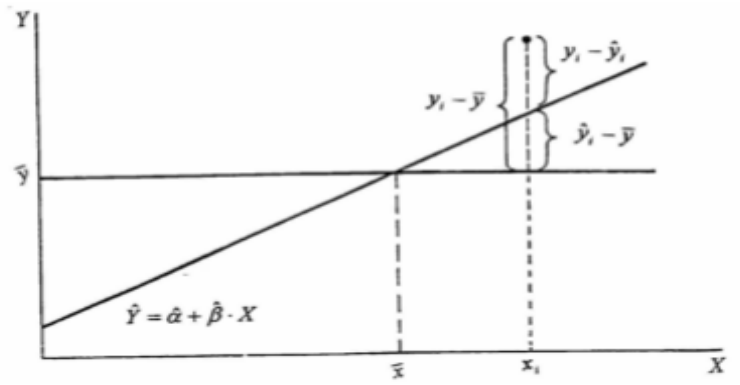
$$\textbf{SSE}=\sum_{i=1}^{\nu}(y_i-\hat{Y}_i)^2\textbf{(16)}$$

Το συνολικό άθροισμα που προκύπτει είναι το εξής: *SSTO=SSR+SSE* και μετράει τη συνολική μεταβλητότητα των παρατηρήσεων y_i. Εκφράζει την αβεβαιότητα κατά την πρόβλεψη της μεταβλητής Y όταν δε χρησιμοποιείται το X. Το SSR μετράει τη μεταβλητότητα η οποία μπορεί να οφείλεται στη μεταβλητή X. Κατά συνέπεια το SSE=SSTO-SSR εκφράζει την υπόλοιπη μεταβλητότητα που δεν ερμηνεύεται από την παλινδρόμηση (Χρήστου, 2007).

Ο συντελεστής προσδιορισμού είναι ο λόγος $\mathbf{r^2 = SSR/\ SSTO}$ **(17)** που εκφράζει το ποσοστό της μεταβλητότητας των y_i που απορροφάται από την παλινδρόμηση και λαμβάνει τιμές που αντιστοιχούν στο διάστημα [0,1]. Στη περίπτωση που τα παρατηρούμενα σημεία βρίσκονται πάνω στην ευθεία ελαχίστων τετραγώνων θα έχουμε $\mathbf{yi} = \hat{\mathbf{Y}}\mathbf{i}$ οπότε r^2=1. Θα έχουμε r^2=0 όταν η κλίση της ευθείας ελαχίστων τετραγώνων είναι μηδέν. Όσο πιο κοντά στη μονάδα βρίσκεται τόσο πιο συνεπής εκτιμητής της απλής παλινδρόμησης είναι η ευθεία των ελαχίστων τετραγώνων. Η τιμή του συντελεστή προσδιορισμού μπορεί να βρίσκεται μεταξύ 0 και 1 στις διάφορες πρακτικές που χρησιμοποιείται.

Παρακάτω παρουσιάζεται ένα διάγραμμα που απεικονίζει όσα προαναφέρθηκαν.

Σχήμα 3.4: Διάγραμμα εκτίμησης της ευθείας παλινδρόμησης



Πηγή: (Χρήστου, 2007)

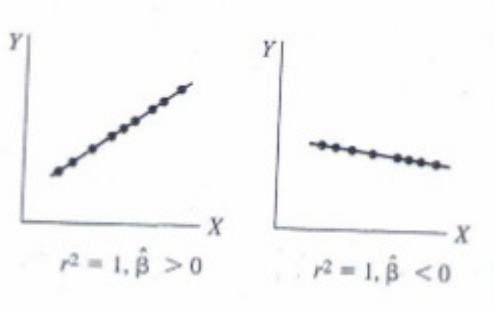
Από το παραπάνω σχήμα διακρίνουμε την ευθεία παλινδρόμησης. Τυπικό σφάλμα εκτίμησης ονομάζεται η μέση απόκλιση μεταξύ της πραγματικής και της εκτιμούμενης τιμής της μεταβλητής, την οποία συμβολίζουμε με s και γράφεται ως εξής:

$$\mathbf{S} = \sqrt{\frac{1}{\nu-2} \cdot \sum_{i=1}^{\nu} (y_i - \hat{y}\mathbf{i})^2} = \sqrt{\frac{SSE}{\nu-2}} \quad \mathbf{(18)}$$

Αποτελεί ένα μέτρο της διασποράς των (x_i, y_i), γύρω από την ευθεία ελαχίστων τετραγώνων (το s^2 είναι μια εκτίμηση της διασποράς των σφαλμάτων).

Η ποσότητα $1-r^2$ εκφράζει το ποσοστό της συνολικής μεταβλητότητας που οφείλεται στο τυχαίο σφάλμα. Τέλος, αξίζει να σημειωθεί ότι r^2 δεν μετρά πόσο μεγάλη είναι η κλίση $\hat{\beta}$ της ευθείας παλινδρόμησης. Αυτό το γεγονός ερμηνεύεται από τα παρακάτω διαγράμματα.

Σχήμα 3.5: Συντελεστής προσδιορισμού R^2



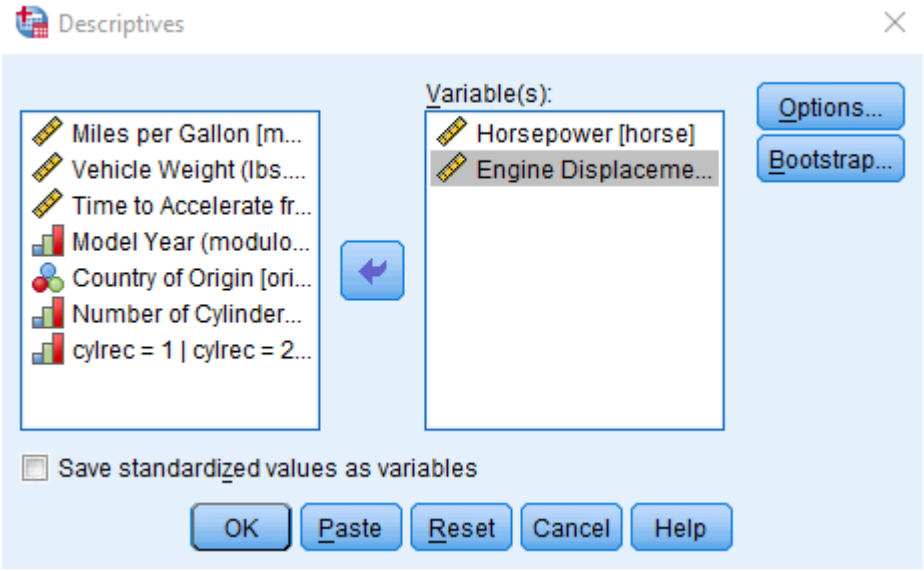
Πηγή: (Χρήστου, 2007)

3.6 Απλή γραμμική παλινδρόμηση με χρήση του SPSS

Το SPSS αποτελεί ένα στατιστικό πρόγραμμα το οποίο χρησιμοποιείται για την εκτέλεση στατιστικών ερευνών. Παρέχει πληροφορίες για τα περιγραφικά στατιστικά χαρακτηριστικά των δεδομένων που εξετάζονται και προσφέρει διαγράμματα με βάση τα οποία ο ερευνητής μπορεί να αντλήσει σημαντικά αποτελέσματα για την έρευνά του. Μέσα από αυτό το στατιστικό εργαλείο μπορούμε να εξάγουμε αποτελέσματα εκτελώντας μια απλή γραμμική παλινδρόμηση.

Τα δεδομένα που χρησιμοποιούμε για να εφαρμόσουμε τη μέθοδο της απλής γραμμικής παλινδρόμησης βρίσκονται στο φάκελο SPSS και αποτελούν δεδομένα που αφορούν αυτοκίνητα (Cars.sav). Οι μεταβλητές που χρησιμοποιούνται είναι τα χιλιόμετρα που διανύει ένα αυτοκίνητο δηλαδή η μέση κατανάλωση, το βάρος του, η ιπποδύναμη του, το μοντέλο, ο κυβισμός του, η χώρα κατασκευής του κα. Το πρώτο βήμα που ακολουθούμε είναι να «περάσουμε» τα δεδομένα στο SPSS. Υπάρχουν 2 φύλλα εργασίας, στο ένα καταχωρούμε τις μεταβλητές που εξετάζουμε και ονομάζεται Variable view και στο άλλο καταχωρούμε τις παρατηρήσεις των μεταβλητών αυτών και ονομάζεται Data view. Στην προκειμένη περίπτωση επειδή τα δεδομένα μας βρίσκονται ήδη στο φάκελο SPSS εκτελούμε εισαγωγή δεδομένων και τα δεδομένα μας εμφανίζονται στα 2 φύλλα εργασίας. Προκειμένου να εξάγουμε τα περιγραφικά χαρακτηριστικά των παραπάνω δεδομένων επιλέγουμε από τη γραμμή εργαλειών Analyze→ Descriptive Statistics→ Descriptives θα εμφανιστεί το παράθυρο όπως φαίνεται στην Εικόνα 3.6. Με τα βελάκια περνάμε δεξιά τις μεταβλητές που θέλουμε να εξάγουμε τα περιγραφικά μέτρα. Στη συγκεκριμένη περίπτωση επιλέγουμε τις μεταβλητές που σχετίζονται με την ιπποδύναμη και τον κυβισμό των αυτοκινήτων. Αν πατήσουμε την επιλογή Options μπορούμε να επιλέξουμε τα στατιστικά περιγραφικά μέτρα που επιθυμούμε να εξετάσουμε. Τα περιγραφικά μέτρα που δίνονται προεπιλεγμένα από το ίδιο το πρόγραμμα SPSS είναι ο μέσος όρος (mean), η τυπική απόκλιση (Std. Deviation), η ελάχιστη (Minimum) και η μέγιστη (Maximum) τιμή. Αν θέλουμε επιλέγουμε και άλλα περιγραφικά μέτρα εκτός των ήδη επιλεγμένων.

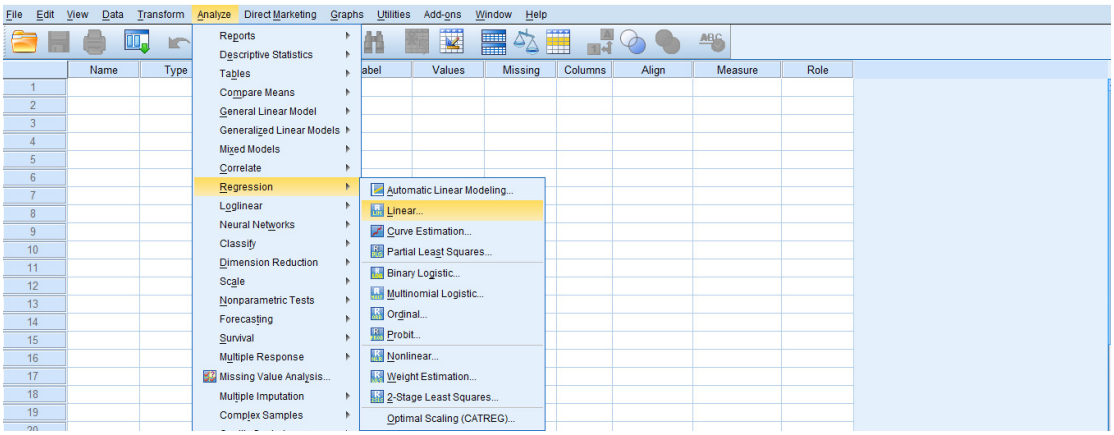
Εικόνα 3.6: Περιγραφικά χαρακτηριστικά δεδομένων



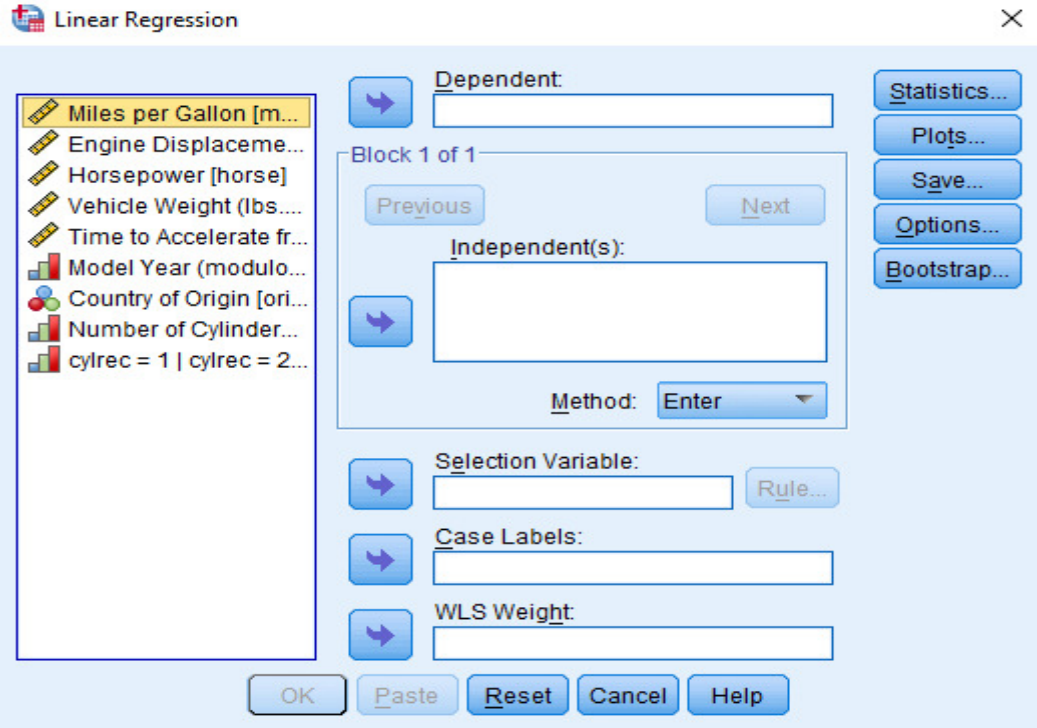
Πηγή: (Τσαγρής, 2008)

Προκειμένου να εκτελέσουμε την απλή γραμμική παλινδρόμηση δηλαδή την ευθεία γραμμικής παλινδρόμησης μαζί με κάποια διαγνωστικά μέτρα μέσω του SPSS επιλέγουμε από το μενού επιλογών τα εξής βήματα: Analyze→Regression→Linear. Αυτό φαίνεται και από τις παρακάτω εικόνες.

Εικόνα 3.7 : Απλή γραμμική παλινδρόμηση στο SPSS

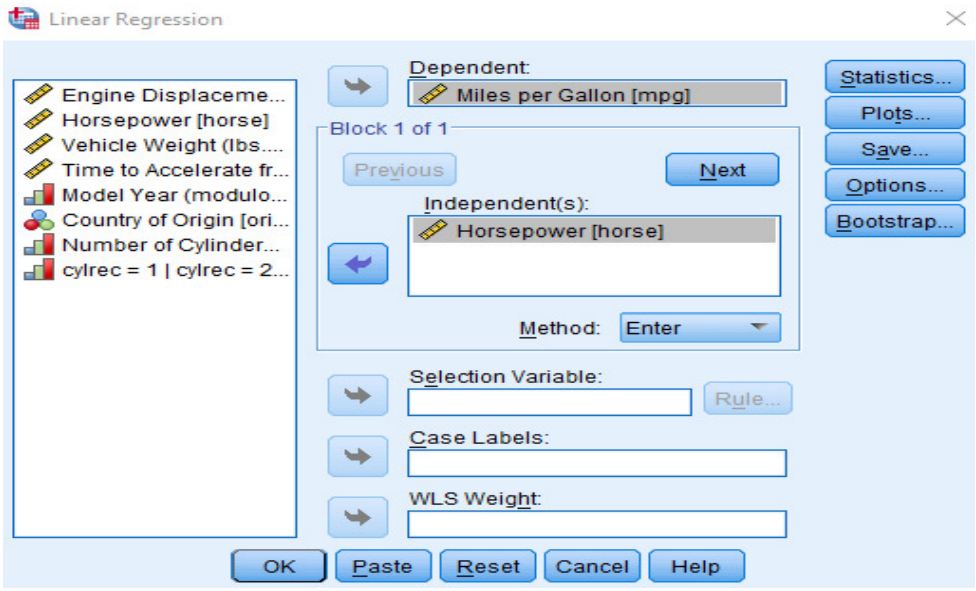


Εικόνα 3.8: Βήματα για την απλή γραμμική παλινδρόμηση στο SPSS



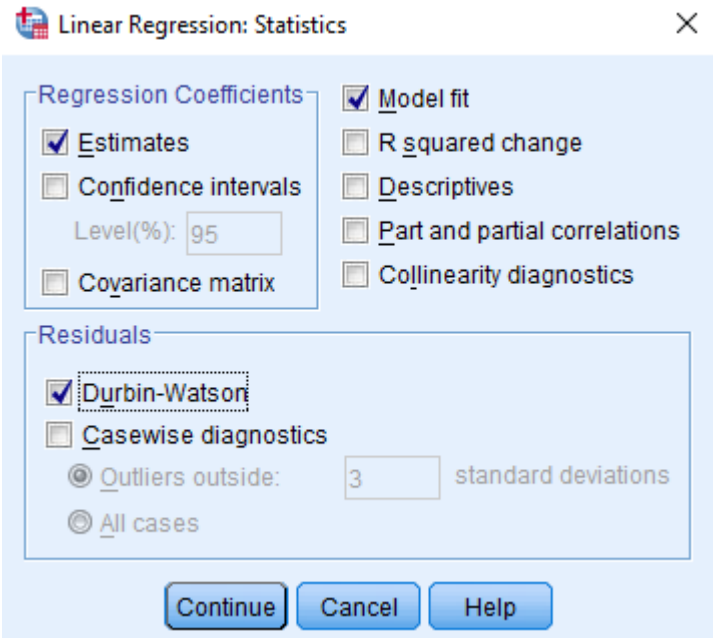
Μετά την ένδειξη Linear εμφανίζονται 2 στήλες, στην αριστερή στήλη εμφανίζονται οι μεταβλητές που έχουμε καταχωρήσει και στα δεξιά εμφανίζονται κουτάκια όπου θα πρέπει να επιλέξουμε πρώτα την εξαρτημένη μεταβλητή που θα χρησιμοποιήσουμε για την εκτίμηση της απλής γραμμικής παλινδρόμησης. Στη συνέχεια στο από κάτω κουτάκι θα πρέπει να επιλέξουμε την ανεξάρτητη μεταβλητή προκειμένου να εκτιμήσουμε την επίδραση που ασκεί η ανεξάρτητη μεταβλητή πάνω στην εξαρτημένη. Επιλέγουμε ως εξαρτημένη μεταβλητή την Miles per Gallon και ως ανεξάρτητη μεταβλητή την Horsepower (ιπποδύναμη). Όλα τα παραπάνω απεικονίζονται στην παρακάτω εικόνα.

Εικόνα 3.9: Επιλογή εξαρτημένης και ανεξάρτητης μεταβλητής στο SPSS



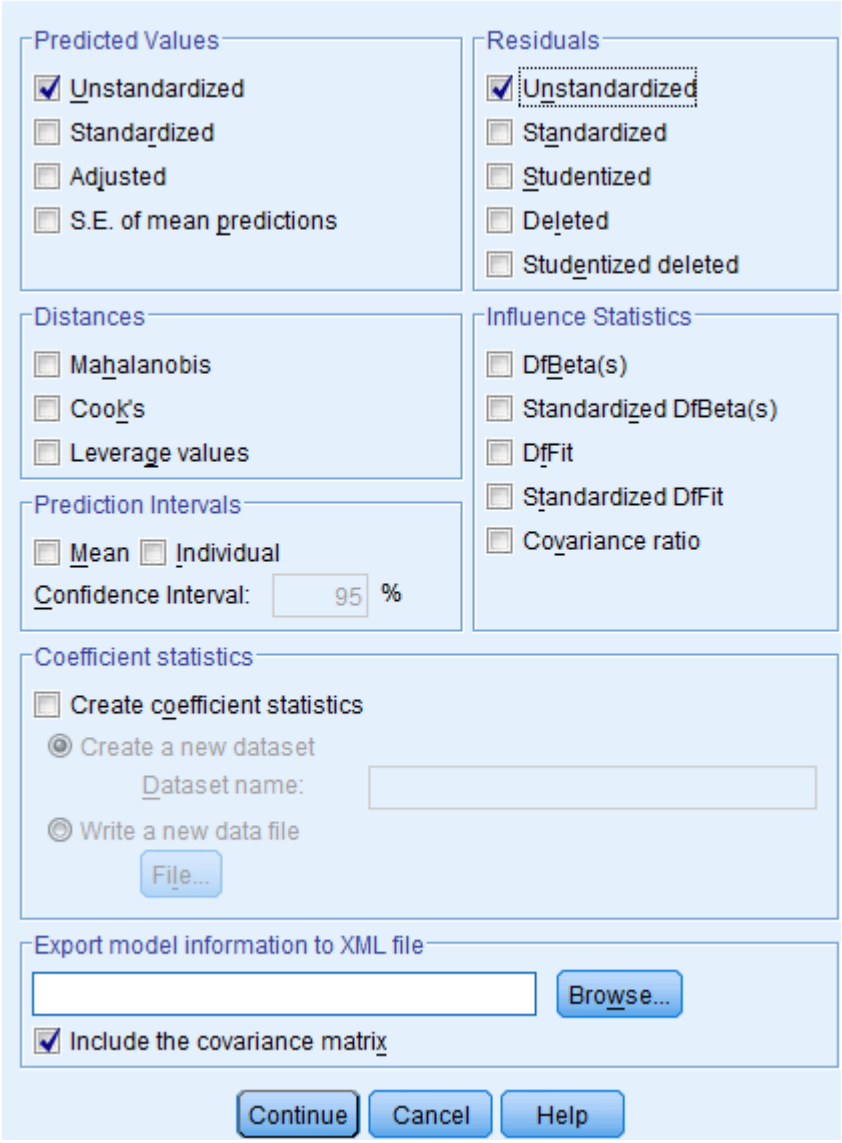
Στην περίπτωση που γνωρίζουμε τη χρονική σειρά με βάση την οποία πραγματοποιήθηκαν οι μετρήσεις μπορούμε να επιλέξουμε από την επιλογή Statistics στο κάτω μέρος, να υπολογιστεί το τεστ των Durbin-Watson το οποίο θέλουμε να χρησιμοποιήσουμε για να ελέγξουμε αν υπάρχει συσχέτιση μεταξύ των καταλοίπων (Εικόνα 3.10). Μέσα από αυτή την επιλογή μας δίνεται η δυνατότητα να επιλέξουμε να εμφανιστούν διαγνωστικά μέτρα συγγραμμικότητας τα οποία θα αναφέρουμε στην επόμενη ενότητα όπου θα μιλήσουμε για την πολλαπλή γραμμική παλινδρόμηση, στην περίπτωση δηλαδή που χρησιμοποιούμε παραπάνω από μια ανεξάρτητες μεταβλητές.

Εικόνα 3.10: Επιλογή τεστ των Durbin-Watson



Στη συνέχεια επιλέγουμε την ένδειξη Continue και εν συνεχεία Save και εμφανίζεται η Εικόνα 3.11 όπου επιλέγουμε να αποθηκεύσουμε τις μη τυποποιημένες εκτιμηθείσες τιμές καθώς και τα μη τυποποιημένα κατάλοιπα. Στην περίπτωση που θέλουμε να κατασκευαστούν διαγράμματα για τον απαραίτητο έλεγχο υποθέσεων όπως θα δούμε σε επόμενη ενότητα, επιλέγουμε την ένδειξη Plots όπως φαίνεται στην Εικόνα 3.9.

Εικόνα 3.11 Save



Προκειμένου να κάνουμε το τεστ κανονικότητας των Kolmogorov-Smirnov, είναι απαραίτητο να αποθηκεύσουμε τα κατάλοιπα. Πατάμε Continue και στη συνέχεια πατώντας OK στο παράθυρο στην Εικόνα 3.9 εμφανίζονται σε πίνακες τα ακόλουθα αποτελέσματα.

Πίνακας 2. Εξαρτημένη: Miles per Gallon- Συντελεστής προσδιορισμού

```
[DataSet1] E:\παλλινδρόμηση\Cars (1).sav
```

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Horsepower[b]	.	Enter

a. Dependent Variable: Miles per Gallon

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,771[a]	,595	,594	4,974	,964

a. Predictors: (Constant), Horsepower

b. Dependent Variable: Miles per Gallon

Η τιμή R αντιπροσωπεύει την απόλυτη τιμή του συντελεστή γραμμικής συσχέτισης. Το R Square όπως αναφέρθηκε σε προηγούμενη ενότητα είναι το τετράγωνο του συντελεστή γραμμικής συσχέτισης δηλαδή ο συντελεστής προσδιορισμού. Ο συντελεστής προσδιορισμού αποδεικνύει το ποσοστό μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από το γραμμικό μοντέλο που προσαρμόσαμε δηλαδή από την επίδραση της ανεξάρτητης μεταβλητής πάνω στην εξαρτημένη. Το συγκεκριμένο μοντέλο του παραδείγματος που χρησιμοποιούμε εξηγεί το 59.5% της μεταβλητότητας των δεδομένων. Δηλαδή το 59.5% της μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής εξηγείται μέσω του μοντέλου από την ανεξάρτητη μεταβλητή. Ο προσαρμοσμένος συντελεστής προσδιορισμού (Adjusted R Square) λαμβάνει υπόψη του και το μέγεθος του δείγματος γι' αυτό ονομάζεται προσαρμοσμένος.

Πίνακας 3: Ανάλυση διακύμανσης

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14169,756	1	14169,756	572,709	,000[b]
	Residual	9649,237	390	24,742		
	Total	23818,993	391			

a. Dependent Variable: Miles per Gallon

b. Predictors: (Constant), Horsepower

Στον Πίνακα 3 παρουσιάζεται η ανάλυση διακύμανσης. Η F στατιστική είναι ένας έλεγχος που βασίζεται στην F κατανομή και είναι υπεύθυνη για τον έλεγχο της

περίπτωσης που όλοι οι παράμετροι του μοντέλου είναι μηδέν ή την περίπτωση που έστω και μια παράμετρος είναι διάφορη του μηδενός. Οι σκιασμένοι αριθμοί στον πίνακα της ανάλυσης διακύμανσης δείχνουν την διακύμανση του μοντέλου συγκεκριμένα ο (14169.756) δείχνει τη διακύμανση που εξηγείται από το μοντέλο που προσαρμόσαμε και ο δεύτερος (23818.993) δείχνει τη συνολική διακύμανση των δεδομένων. Η διαφορά των δυο αυτών αριθμών δείχνει τη διακύμανση που δεν εξηγείται από το μοντέλο. Το πηλίκο αυτών των δύο αριθμών αποτελεί το συντελεστή προσδιορισμού.

Το μοντέλο της απλής γραμμικής παλινδρόμησης που εκτιμάμε ή αλλιώς η ευθεία ελαχίστων τετραγώνων, όπως έχει ήδη αναφερθεί είναι της μορφής $y=\alpha+ \beta x + e_i$, όπου y είναι η εξαρτημένη μεταβλητή, x η ανεξάρτητη μεταβλητή και α, β οι παράμετροι του μοντέλου τις οποίες εκτιμάμε. Ο όρος e_i είναι το κατάλοιπο της i-οστής τιμής. Μέσα από το στατιστικό πρόγραμμα SPSS παίρνουμε τα αποτελέσματα για την απλή γραμμική παλινδρόμηση που εκτιμήσαμε και αυτή φαίνεται στον πίνακα 4. Είναι της μορφής:

Miles per Gallon=39.855-0.157*Horsepower

Όπου η εξαρτημένη μεταβλητή είναι η Miles per Gallon και η ανεξάρτητη μεταβλητή η Horsepower. Η τιμή 39.855 είναι ο σταθερός όρος της εκτίμησης παλινδρόμησης δηλαδή η παράμετρος α και η τιμή 0.157 είναι ο συντελεστής της ανεξάρτητης μεταβλητής δηλαδή η παράμετρος β.

Πίνακας 4: Εκτίμηση παραμέτρων

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	39,855	,730		54,578	,000
	Horsepower	-,157	,007	-,771	-23,931	,000

a. Dependent Variable: Miles per Gallon

Πιο συγκεκριμένα η σκιασμένη στον Πίνακα 4 τιμή 39.855 δηλαδή ο σταθερός όρος είναι η τιμή στην οποία η ευθεία ελαχίστων τετραγώνων τέμνει το κάθετο άξονα των y'y. Η τιμή -0.157 αποτελεί τη κλίση της ευθείας. Επίσης φανερώνει την επίδραση της ανεξάρτητης μεταβλητής στην εξαρτημένη. Δηλαδή για κάθε αύξηση της ανεξάρτητης μεταβλητής κατά 1 μονάδα η εκτιμώμενη μέση τιμή της εξαρτημένης μεταβλητής μειώνεται ή αυξάνεται κατά β μονάδες. Στην περίπτωση που εξετάζουμε, για μία αύξηση της ιπποδύναμης κατά 1 μονάδα, παρατηρείται μείωση της εκτιμώμενης μέσης κατανάλωσης κατά 0,157 μονάδες. Με άλλα λόγια καθώς αυξάνεται η ιπποδύναμη κατά 1%, η εκτιμώμενη μέση κατανάλωση μειώνεται κατά 15,7%. Η τελευταία στήλη του πίνακα που είναι σκιασμένη περιλαμβάνει τα παρατηρηθέντα επίπεδα στατιστικής σημαντικότητας, τα οποία χρησιμοποιούνται προκειμένου να εξάγουμε συμπεράσματα σχετικά με τη στατιστική σημαντικότητα των παραμέτρων α και β του μοντέλου που εξετάζουμε. Δηλαδή διαπιστώνουμε αν η παραπάνω συσχέτιση της εξαρτημένης μεταβλητής με την ανεξάρτητη είναι στατιστικά σημαντική. Οι υποθέσεις που ελέγχονται εδώ όσον αφορά στους συντελεστές α και β είναι οι εξής: Ho: $\alpha=0$ H1: $\alpha\neq0$ και Ho: $\beta=0$ H1: $\beta\neq0$.

Στην περίπτωση που εξετάζουμε παρατηρούμε ότι και οι δύο p-value είναι μικρότερες του 0.05.[2] Συνεπώς καταλήγουμε στο συμπέρασμα ότι και οι δύο μηδενικές υποθέσεις απορρίπτονται. Δηλαδή, οι δύο συντελεστές είναι στατιστικά σημαντικοί γεγονός που αποδεικνύει την σημαντικότητά τους για το εξεταζόμενο μοντέλο. Η τιμή p-value καθορίζει την αξιοπιστία της ανεξάρτητης μεταβλητής x για την πρόβλεψη της εξαρτημένης μεταβλητής y. Σε επίπεδο στατιστικής σημαντικότητας α=0,05, όταν το p-value είναι μικρότερο του 0,05 συμπεραίνουμε ότι η σχέση του x και του y είναι στατιστικά σημαντική.

Πίνακας 5: Περιγραφικά μέτρα καταλοίπων

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3,64	32,61	23,45	6,020	392
Residual	-16,212	16,980	,000	4,968	392
Std. Predicted Value	-3,290	1,523	,000	1,000	392
Std. Residual	-3,259	3,414	,000	,999	392

a. Dependent Variable: Miles per Gallon

Ο Πίνακας 5 περιλαμβάνει τα περιγραφικά στατιστικά χαρακτηριστικά των καταλοίπων. Η μέση τιμή των καταλοίπων είναι ίση με το 0. Η πρώτη στήλη του πίνακα δείχνει την ελάχιστη τιμή των καταλοίπων, η δεύτερη στήλη την μέγιστη τιμή τους, η τρίτη την μέση τιμή, η τέταρτη την τυπική απόκλιση και τέλος στη τελευταία στήλη φαίνεται ο αριθμός των παρατηρήσεων N του δείγματος μας. Φαίνεται δηλαδή ότι το δείγμα που χρησιμοποιείται στην προκειμένη περίπτωση είναι 392 παρατηρήσεις. Μια από τις υποθέσεις που έχουν τεθεί κατά την εκτίμηση της απλής γραμμικής παλινδρόμησης όσον αφορά τα κατάλοιπα είναι ότι ακολουθούν την κανονική κατανομή με μέσο το 0. Η ικανοποίηση της κανονικότητας των καταλοίπων μπορεί να ελεγχθεί είτε μέσω διαγραμμάτων (P-P ή Q-Q Plots) είτε μέσω του τεστ των Kolmogorov Smirnov. Στην περίπτωση που δεν μπορούμε να υποθέσουμε την κανονικότητα των καταλοίπων μετασχηματίζουμε τις εξαρτημένες μεταβλητές ώστε να έχουμε ως αποτέλεσμα την κανονικότητα των καταλοίπων. Στην Εικόνα 3.11 είχαμε επιλέξει την αποθήκευση των καταλοίπων και των εκτιμηθεισών τιμών για την εκτίμηση της ευθείας ελαχίστων τετραγώνων (Τσαγρής, 2008).

Προκειμένου να ελεγχθεί η ανεξαρτησία και η ομοσκεδαστικότητα των καταλοίπων εξάγουμε ένα διάγραμμα διασποράς το οποίο περιέχει τις εκτιμηθείσες τιμές στον οριζόντιο άξονα και τα κατάλοιπα στο κάθετο άξονα. Στην περίπτωση που τα κατάλοιπα είναι ανεξάρτητα θα εμφανιστεί ένα "σύννεφο" σημείων στο διάγραμμα και όχι ένα σχήμα. Ένας άλλος τρόπος να ελεγχθεί η ομοσκεδαστικότητα των καταλοίπων είναι ο υπολογισμός της συνδιακύμανσης των καταλοίπων και των εκτιμηθέντων τιμών. Στην περίπτωση που παρατηρείται ανεξαρτησία τότε η συνδιακύμανση θα ισούται με το μηδέν. Το αντίστροφο δεν ισχύει πάντα. Αν όμως θεωρήσουμε ότι δεν ικανοποιείται η υπόθεση της ανεξαρτησίας των καταλοίπων τότε δεν εφαρμόζουμε σαν μέθοδο τη γραμμική παλινδρόμηση αλλά χρησιμοποιούμε διαφορετικές τεχνικές. Στην περίπτωση που τα κατάλοιπα είναι σειριακά

[2] Οι τιμές P value ελέγχουν την υπόθεση ότι κάθε συντελεστής είναι διαφορετικός του μηδενός. Για να συμβεί αυτό η τιμή P value πρέπει να είναι μικρότερη του 0,05.

συσχετισμένα χρησιμοποιούμε μονότονη παλινδρόμηση η οποία είναι εφικτή από το SPSS (Τσαγρής, 2008).

Αν η υπόθεση της ομοσκεδαστικότητας ικανοποιείται είναι σημαντικό τα σημεία του διαγράμματος στο κατακόρυφο άξονα σε όλο το εύρος τους να είναι σταθερά καθώς μεταβαίνουμε στον οριζόντιο άξονα. Στη περίπτωση που το εύρος των σημείων αυξάνεται ή μειώνεται καθώς μεταβαίνουμε στα δεξιά του οριζόντιου άξονα τότε δεν είναι δυνατό να υποθέσουμε ομοσκεδαστικότητα των καταλοίπων. Η αντιμετώπιση αυτού του προβλήματος γίνεται μετασχηματίζοντας τις ανεξάρτητες ή και τις εξαρτημένες μεταβλητές. Στην περίπτωση που δεν έχουμε επίλυση του προβλήματος θα πρέπει να εφαρμοστούν μη παραμετρικές μέθοδοι παλινδρόμησης (Theil, παλινδρόμηση στις τάξεις μεγέθους). Αξίζει να σημειωθεί ότι ακόμα και σε αυτές τις περιπτώσεις η ετεροσκεδαστικότητα αποτελεί πρόβλημα. Ένας άλλος τρόπος αντιμετώπισης της ετεροσκεδαστικότητας είναι η χρησιμοποίηση γενικευμένων γραμμικών μοντέλων. Σε περίπτωση που δεν ικανοποιείται η υπόθεση της ομοσκεδαστικότητας εφαρμόζουμε τον λογαριθμικό μετασχηματισμό στις τιμές της ανεξάρτητης μεταβλητής και εκτιμούμε την εξίσωση της ευθείας παλινδρόμησης πάνω στη μετασχηματισμένη μεταβλητή (Τσαγρής, 2008).

Κεφάλαιο 4: Πολλαπλή γραμμική παλινδρόμηση

4.1 Βασικές έννοιες

Στη στατιστική και γενικότερα στις στατικές εφαρμογές παρατηρείται το πρόβλημα της μελέτης της σχέσης ανάμεσα σε δυο ή περισσότερες τυχαίες μεταβλητές. Όταν έχουμε να μελετήσουμε περισσότερες από μια τυχαίες μεταβλητές εφαρμόζουμε τη μέθοδο της πολλαπλής παλινδρόμησης. Για παράδειγμα όταν θέλουμε να εξάγουμε συμπεράσματα για τη σχέση μεταξύ του ύψους μιας ομάδας ατόμων και του εισοδήματός τους, της ηλικίας τους, της κατανάλωσης των ατόμων στην εταιρία που εργάζονται, της εργασιακής τους εμπειρίας κλπ (Ιωαννίδης, 2005).

Το πρόβλημα που γεννάται είναι η διερεύνηση της σχέσης που συνδέει τις εξεταζόμενες μεταβλητές, αν υφίσταται κάποια σχέση και εν συνεχεία να προσδιορίσουμε αυτή τη σχέση βασιζόμενοι σε κάποιες παρατηρήσεις. Η μελέτη αυτή είναι απόλυτα σημαντική διότι τα αποτελέσματα που εξάγονται χρησιμοποιούνται πολύ συχνά για προβλέψεις. Για παράδειγμα η πρόβλεψη του ποσοστού ανεργίας σε μακροοικονομικό επίπεδο με βάση τα επιτόκια της αγοράς, την μεταβολή του Α.Ε.Π. κλπ. Ιδιαίτερα στις ιδιωτικές εταιρίες απαιτείται η πρόβλεψη μεταβλητών όπως είναι η ζήτηση των προϊόντων, τα επιτόκια, η μεταβολή του ρυθμού πληθωρισμού, οι τιμές των πρώτων υλών, το εργατικό κόστος κλπ (Gerald, 2010).

Η Πολλαπλή Παλινδρόμηση αποσκοπεί στη κατασκευή ενός μοντέλου που να εξηγεί σε ικανοποιητικό βαθμό τη σχέση που συνδέει μια εξαρτημένη συνεχής μεταβλητή Y με μία ή περισσότερες συνεχείς ανεξάρτητες μεταβλητές X1 ,X2,…, X p. Στην περίπτωση που η σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών είναι γραμμική τότε η σχέση αυτή περιγράφεται με βάση ένα γραμμικό μοντέλο. Αυτή η σχέση σύνδεσης έχει την εξής μορφή:

$$\mathbf{Y = b_0 + b_1X_1 + b_2X_2 + ... + b_PX_P} \quad \mathbf{(4.1)}$$

Αυτές οι μεταβλητές δεν είναι γνωστές από πριν, επιλέγονται οι πιο κατάλληλες από ένα σύνολο μεταβλητών που έχουμε διαθέσιμες. Εφαρμόζοντας τη τεχνική της Πολλαπλής Γραμμικής Παλινδρόμησης (ΠΓΠ) εντοπίζουμε το μοντέλο που είναι το πιο ικανοποιητικό για τη μελέτη μας ή διαπιστώνουμε ότι δεν υπάρχει κανένα ικανοποιητικό. Αποτελεί περίπλοκη διαδικασία να δημιουργήσουμε μια μαθηματική εξίσωση για να περιγράψουμε ένα φαινόμενο όταν δεν γνωρίζουμε τη σχέση που υπάρχει μεταξύ των μεταβλητών. Μέσω της διαδικασίας προσδιορισμού του μοντέλου της Γραμμικής Παλινδρόμησης επιδιώκουμε να περιγράψουμε με το πιο βέλτιστο τρόπο την πληροφορία που μας δίνεται από τα δεδομένα που χρησιμοποιούμε (Κιντής, 2010).

Όμως στη πράξη δεν μπορεί να περιγράψει κανένα μοντέλο και κατά συνέπεια γραμμικό μοντέλο με ακρίβεια τις πληροφορίες που εξετάζουμε από ένα σύνολο δεδομένων. Όσο καλά και να εφαρμόζεται η γραμμή της πολλαπλής παλινδρόμησης πάνω στα δεδομένα που εξετάζουμε, θα υπάρχει πάντα ένα μέρος της πληροφορίας που δε θα μπορεί να ερμηνευτεί μέσω του μοντέλου. Αυτό που δεν μπορεί να ερμηνευτεί από το γραμμικό μοντέλο ονομάζεται λάθος της παλινδρόμησης και δεν εμπεριέχει καμία συστηματική σχέση. Έτσι συμπεριλαμβάνοντας και το παράγοντα λάθος στο μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης έχουμε την παρακάτω εξίσωση (Χάλκος, 2011):

$$\mathbf{Y=b_0 + b_1X_1 +b_2X_2 + ... + b_p X_P +e \ (4.2)}$$

όπου:

- Y, εξαρτημένη μεταβλητή
- $X_1, X_2 ..., X_P$ ανεξάρτητες μεταβλητές
- b_0 ο σταθερός όρος, $b_1, b_2, \ldots, b_p$ οι συντελεστές κλίσης κάθε μιας ανεξάρτητης μεταβλητής
- e ο παράγοντας λάθους δηλαδή τα κατάλοιπα

Είναι αναγκαίο πριν προχωρήσουμε στην εφαρμογή της ΠΓΠ να μελετήσουμε με αναλυτικό τρόπο τα δεδομένα μας στα οποία θα εφαρμόσουμε την τεχνική, κατανοώντας με αυτό τον τρόπο το είδος των μεταβλητών, εξαρτημένης και ανεξάρτητων και να διερευνήσουμε την ποιότητα των δεδομένων, να ελέγξουμε αν υπάρχουν τιμές που λείπουν (missing values) καθώς και διάφορα άλλα θέματα που προκύπτουν από την ανάλυση των δεδομένων (Gerald, 2010).

Κατά την εφαρμογή της ΠΓΠ δημιουργούνται κάποιοι περιορισμοί. Η τεχνική δεν είναι δυνατό να εφαρμοστεί σε όλα τα δεδομένα. Μελετά τη σχέση μεταξύ μιας εξαρτημένης συνεχούς μεταβλητής και μίας ή περισσότερων συνεχών ανεξάρτητων μεταβλητών. Συνεπώς ο πρώτος περιορισμός που δημιουργείται είναι το είδος των μεταβλητών που θα χρησιμοποιήσουμε από ένα σύνολο δεδομένων. Υπάρχουν κάποιες αναγκαίες συνθήκες που θα πρέπει να ισχύουν (Χρήστου, 2007).

Πρώτον, η εξαρτημένη μεταβλητή θα πρέπει να είναι συνεχής ποσοτική, κλίμακας (Scale) ακολουθώντας τη Κανονική Κατανομή με σταθερή διακύμανση. Δεύτερον όσον αφορά τις ανεξάρτητες μεταβλητές θα πρέπει να είναι είτε συνεχείς είτε κατηγορικές δηλαδή να χωρίζονται σε κατηγορίες. Στη τελευταία περίπτωση εισάγουμε τις κατηγορικές μεταβλητές με τη μορφή ψευδομεταβλητών (dummies)[3]. Τέταρτον είναι αναγκαίο η καθεμία από τις ανεξάρτητες μεταβλητές να συνδέεται με γραμμικό τρόπο με την εξαρτημένη μεταβλητή. Πέμπτο, οι ανεξάρτητες μεταβλητές θα πρέπει να παρουσιάζουν ισχυρή συσχέτιση με την εξαρτημένη μεταβλητή αλλά όχι μεταξύ τους. Τέλος, τα λάθη πρέπει να είναι τυχαία ακολουθώντας σταθερή διακύμανση. Αυτές οι υποθέσεις είναι απαραίτητο να εξετάζονται πριν καθώς και μετά την εφαρμογή της τεχνικής ΠΓΠ και της κατασκευής της. Αν δεν ικανοποιούνται οι παραπάνω υποθέσεις οδηγούμαστε στη μη αξιοπιστία της τεχνικής και σε αμφισβητήσιμα αποτελέσματα (Χάλκος, 2011).

Κατά την εφαρμογή της ΠΓΠ είναι αναγκαίο να κατανοηθεί από τον μελετητή ποια είναι η φύση του προβλήματος, τι θέλει να εξετάσει και να καταλήξει μέσα από τη μελέτη του αν τα αποτελέσματα αυτής της τεχνικής απαντούν στα ερωτήματα που έχει θέσει. Η τεχνική ενδείκνυται στην περίπτωση που ο μελετητής ενδιαφέρεται να διαπιστώσει στατιστικά σημαντικές σχέσεις μεταξύ της εξαρτημένης μεταβλητής και πολλών ανεξάρτητων μεταβλητών προσδιορίζοντας με ακριβή τρόπο πως συνδέονται οι ανεξάρτητες με τις εξαρτημένες μεταβλητές της ανάλυσης. Επιπλέον για την εξαγωγή προβλέψεων η τεχνική ΠΓΠ αποφέρει πολύ καλά αποτελέσματα (Κιντής, 2010).

Για παράδειγμα σε ένα διαιτολόγιο θα πρέπει να τεθούν ερωτήματα που σχετίζονται με τους συνδυασμούς τροφών που παράγουν χοληστερίνη, την πρόβλεψη

[3] Θα γίνει αναφορά παρακάτω

της ποσότητας χοληστερίνης που θα παραχθεί στο αίμα κλπ. Αυτό μπορεί να γίνει διερευνώντας τα δεδομένα που έχουμε διαθέσιμα ή πρέπει να συλλέξουμε ώστε να δώσουμε απαντήσεις στα παραπάνω ερευνητικά ερωτήματα που θέτουμε κατά την εφαρμογή της ΠΓΠ. Είναι αναγκαίο η εξαρτημένη μεταβλητή του δείγματος που χρησιμοποιούμε να είναι ποσοτική δηλαδή συνεχής, οι ανεξάρτητες μεταβλητές να είναι επίσης συνεχείς ή κατηγορικές και να συνάδουν νοηματικά με την επεξήγηση του εξαρτημένου μεγέθους που εξετάζουμε (Χάλκος, 2011).

Πολλές φορές δεν είναι εφικτό να χρησιμοποιηθεί όλο το σύνολο των δεδομένων κατά τη μελέτη της ΠΓΠ. Στη συνέχεια είναι απαραίτητο να διερευνήσουμε αν οι υποθέσεις που αναφέραμε προηγουμένως ικανοποιούνται σε μεγάλο βαθμό προκειμένου να προχωρήσουμε στο τελικό στάδιο εφαρμογής της τεχνικής. Σε πρώτη φάση θα πρέπει να μελετηθεί ποια Κατανομή ακολουθεί η εξαρτημένη μεταβλητή και αν ακολουθεί την κανονική Κατανομή. Θα πρέπει η διακύμανση της να είναι σταθερή. Σε δεύτερη φάση, είναι αναγκαίο να μελετηθεί η σχέση μεταξύ ανεξάρτητων μεταβλητών και εξαρτημένου μεγέθους. Η σχέση μεταξύ κάθε μιας ξεχωριστά ανεξάρτητης μεταβλητής με την εξαρτημένη θα πρέπει να είναι ισχυρή και γραμμική (Χρήστου, 2007).

Στο παράδειγμα του διαιτολογίου η χοληστερίνη θα πρέπει να παρουσιάζει έντονη γραμμική σχέση με κάθε μια από τις ανεξάρτητες μεταβλητές. Οι ανεξάρτητες μεταβλητές από την άλλη μεριά δε θα πρέπει να συσχετίζονται μεταξύ τους. Ο λόγος που το επιθυμούμε αυτό είναι ότι οι ανεξάρτητες μεταβλητές μπορούν να μεταφέρουν διαφορετικές πληροφορίες ως προς την κίνηση των τιμών της εξαρτημένης μεταβλητής γεγονός μας οδηγεί στην περιγραφή μεγαλύτερου μέρους των πληροφοριών που έχουμε για την εξαρτημένη μεταβλητή Y. Στην καλύτερη περίπτωση θα θέλαμε μια ισχυρή συσχέτιση της εξαρτημένης με κάθε μία από τις ανεξάρτητες και μια πολύ μικρή συσχέτιση των ανεξάρτητων μεταβλητών μεταξύ τους (Ζαχαροπούλου, 2010).

Εφόσον έχουν ελεγχθεί οι παραπάνω υποθέσεις και έχουν μελετηθεί διεξοδικά οι σχέσεις των δεδομένων μας προχωράμε στην εφαρμογή της τεχνικής της ΠΓΠ κατασκευάζοντας το μοντέλο που περιγράφει καλύτερα τις σχέσεις που κρύβουν τα δεδομένα που εξετάζουμε. Το μοντέλο που κατασκευάζεται στη τελική μορφή του αποτελεί μια γενίκευση της πληροφορίας του δείγματος που χρησιμοποιούμε και περιγράφει τον τρόπο κίνησης και σύνδεσης των δεδομένων. Η παρατήρηση της βάσης των δεδομένων δεν αρκεί για να εξάγουμε το μοντέλο. Πρέπει πάντα να ορίζουμε την εξαρτημένη μεταβλητή και τις ανεξάρτητες μεταβλητές οι οποίες θα βρίσκονται στο δεξιό μέρος της μαθηματικής συνάρτησης όπως δείξαμε παραπάνω (Champkin, 2013).

Αφού προσδιορίσουμε μαθηματικά το μοντέλο μας παρατηρούμε τη σχέση σύνδεσης των στατιστικά σημαντικών ανεξάρτητων μεταβλητών με την εξαρτημένη. Δηλαδή διερευνάμε κατά πόσο οι ανεξάρτητες μεταβλητές επιδρούν πάνω στην εξαρτημένη και τον βαθμό στον οποίο αυτή η επίδραση είναι στατιστικά σημαντική και άρα θα τη λάβουμε υπόψη μας. Εξάγουμε συμπεράσματα που αφορούν τον τρόπο σύνδεσης μεταξύ των ανεξάρτητων και της εξαρτημένης αλλά και την στατιστική σημαντικότητα της σχέσης που τις συνδέει. Για παράδειγμα συμπεραίνουμε αν η αύξηση της τιμής των αναψυκτικών αυξάνει ή μειώνει την τιμή της χοληστερίνης (Χρήστου, 2007).

Μετά την κατασκευή του μοντέλου της ΠΓΠ είναι αναγκαίος ο έλεγχος της συμπεριφοράς των καταλοίπων δηλαδή των λαθών της μεθόδου όπως αναφέραμε παραπάνω. Μέσω του μοντέλου προσδιορίζουμε μια τιμή πρόβλεψης για την εξαρτημένη μεταβλητή. Η διαφορά που προκύπτει μεταξύ της πραγματικής τιμής της εξαρτημένης από την τιμή αυτή που παίρνει μέσω πρόβλεψης ονομάζεται σφάλμα ή κατάλοιπο της εξίσωσης παλινδρόμησης. Όπως προαναφέρθηκε είναι απαραίτητο τα σφάλματα να ακολουθούν την κανονική κατανομή, να έχουν σταθερή διακύμανση και να είναι μεταξύ τους ανεξάρτητα. Η τελευταία υπόθεση της ανεξαρτησίας ελέγχεται αφού κατασκευαστεί το μοντέλο της παλινδρόμησης. Η μη ύπαρξη του συστηματικού παράγοντα που να συμμετέχει στον τρόπο κίνησης της εξαρτημένης μεταβλητής και στην ανάλυση σαν ανεξάρτητη μεταβλητή επιβεβαιώνεται από την τυχαιότητα των καταλοίπων (Ιωαννίδης, 2005).

Συνοψίζοντας υπάρχει μια σειρά από βήματα που ακολουθούνται κατά την εφαρμογή της ΠΓΠ, αναλύονται διεξοδικά στο παρόν κεφάλαιο και αυτά είναι τα εξής:

- Ορίζουμε τις μεταβλητές, εξαρτημένη και τις ανεξάρτητες
- Υπολογίζουμε τους συντελεστές συσχέτισης ανάμεσα στην εξαρτημένη και στις ανεξάρτητες μεταβλητές
- Ελέγχουμε ως προς τη στατιστική σημαντικότητα για κάθε έναν συντελεστή συσχέτισης – χρησιμοποιούμε τους ελέγχους t
- Υπολογίζουμε την ευθεία ελαχίστων τετραγώνων
- Ελέγχουμε ως προς τη σημαντικότητα τους συντελεστές κλίσης για κάθε μία από τις ανεξάρτητες μεταβλητές, χρησιμοποιούμε τους ελέγχους t
- Ελέγχουμε ως προς τη σημαντικότητα το μοντέλο Πολλαπλής Γραμμικής Παλινδρόμησης χρησιμοποιώντας τον έλεγχο F
- Υπολογίζουμε και ερμηνεύουμε το Συντελεστή Προσδιορισμού
- Υπολογίζουμε τα κατάλοιπα
- Ελέγχουμε ως προς την κανονικότητα τα κατάλοιπα
- Ελέγχουμε ως προς την αυτοσυσχέτιση τα κατάλοιπα
- Ελέγχουμε για πολυσυγγραμικότητα και αντοχή (Gerald, 2010)

4.2 Εξίσωση ελαχίστων τετραγώνων

Όπως στην απλή γραμμική παλινδρόμηση έτσι και στην πολλαπλή ψάχνουμε μια ευθεία με τον καταλληλότερο σταθερό όρο και τους πιο κατάλληλους συντελεστές κλίσης για κάθε ανεξάρτητη μεταβλητή του δείγματός μας έτσι ώστε η ευθεία παλινδρόμησης να περνάει όσο πιο κοντά γίνεται από τις διασπαρμένες παρατηρήσεις. Στην ουσία αναζητούμε την ευθεία που θα ελαχιστοποιεί το τετράγωνο των αποστάσεων των παρατηρήσεων από την ευθεία. Οι αποστάσεις αυτών των παρατηρήσεων από την ευθεία ονομάζονται όπως αναφέραμε και προηγουμένως, κατάλοιπα ή σφάλματα. Κατά συνέπεια η ευθεία υπολογίζεται με τρόπο ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των καταλοίπων (Ζαχαροπούλου, 2010).

Αντιλαμβανόμαστε ότι η ευθεία της πολλαπλής γραμμικής παλινδρόμησης υπολογίζεται μέσω της εφαρμογής της μεθόδου των ελαχίστων τετραγώνων όπως συμβαίνει και στην απλή γραμμική παλινδρόμηση μόνο που εδώ έχουμε περισσότερες από μια ανεξάρτητες μεταβλητές. Η σχέση (4.2) μας δείχνει την ευθεία

ελαχίστων τετραγώνων για μια πολλαπλή γραμμική παλινδρόμηση με περισσότερους από έναν συντελεστές κλίσης (β_0 σταθερός όρος, β_1, $\beta_{2...,}\beta p$ συντελεστές κλίσης).

Ο σταθερός όρος β_0 είναι η E(Y) για $X_1 = X_2 = ... = X_p = 0$ ενώ το βi (i= 1,2,...,p) αποδεικνύει τη μεταβολή της E(Y) σε μια αύξηση της μεταβλητής X_i κατά μια μονάδα υπό την προϋπόθεση ότι όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

Η συνάρτηση παλινδρόμησης (regression function) ή αλλιώς συνάρτηση ανταπόκρισης (response function) του μοντέλου (4.2) είναι

$$\mathbf{E\,(Y \mid x_1\,,\, x_2,\ldots,x_p\,) = \alpha + \beta_1\, X_1 + \beta_2\, X_2 + ...+ \beta_p\, X_p\,(4.3)}$$

Η εκτιμήτρια της επιφάνειας παλινδρόμησης θα είναι:

$$\hat{Y} = \boldsymbol{\hat{\alpha} + \hat{\beta}_1\, X_1 + \hat{\beta}_2\, X_2 + \ldots + \hat{\beta}_p\, X_p}\ \mathbf{(4.4)}$$

Οι εκτιμώμενοι συντελεστές κλίσης προκύπτουν από τη μέθοδο των ελαχίστων τετραγώνων. Εκτός από την προσαρμογή με τη μέθοδο των ελαχίστων τετραγώνων του μοντέλου που εξετάζουμε σε μια σειρά δεδομένων προκύπτουν κάποια προβλήματα. Στην πολλαπλή παλινδρόμηση δημιουργείται το πρόβλημα κατά πόσον μερικοί από τους όρους $\beta_i X_i$ στο μοντέλο συνεισφέρουν σημαντικά στο να ερμηνεύσουν τη διακύμανση που παρατηρείται στην εξαρτημένη μεταβλητή Y_i. Η πολλαπλή παλινδρόμηση δίνει τη δυνατότητα της στατιστικής συμπερασματολογίας για να καθοριστεί ο βαθμός στον οποίο μια μεταβλητή είναι σημαντική μέσω του ελέγχου της μηδενικής υπόθεσης $H_0 : \beta_i = 0$ έναντι της εναλλακτικής $H_1 : \beta i \neq 0$, i =1,2,...,p. Αν η H_0 δεν απορριφθεί για κάποια τιμή του i καταλήγουμε ότι δεν υπάρχουν ικανά στοιχεία ώστε να μας πείσουν ότι η αντίστοιχη μεταβλητή έχει σημαντική συνεισφορά στο μοντέλο. Στην περίπτωση αυτή ο όρος $\beta_i X_i$ διαγράφεται από το μοντέλο και έτσι η διαδικασία απλοποιείται (Χρήστου, 2007).

Αποτελεί πολύ σημαντικό ζήτημα τόσο στην απλή γραμμική παλινδρόμηση όσο και στην πολλαπλή γραμμική παλινδρόμηση να αποδεικνύεται αν υπάρχει στατιστικά σημαντική γραμμική σχέση σε κάθε ανεξάρτητη μεταβλητή με την εξαρτημένη. Αυτό μπορεί να αποδειχθεί με τη χρήση του ελέγχου t. Όπως είδαμε παραπάνω ο t έλεγχος κάνει μια μηδενική υπόθεση ότι ο συντελεστής συσχέτισης ισούται με μηδέν ενώ στην εναλλακτική υπόθεση θέτει τον συντελεστή συσχέτισης διάφορο του μηδενός. Η τιμή του ελέγχου t τίθεται σε σύγκριση ως τιμή με το κρίσιμο σημείο της κατανομής Τ για συγκεκριμένο επίπεδο εμπιστοσύνης, για το μέγεθος δείγματος που χρησιμοποιούμε. Στη περίπτωση που η τιμή t είναι μεγαλύτερη από τη κρίσιμη τιμή τότε απορρίπτεται η πρόταση της μηδενικής υπόθεσης. Σε αντίθετη περίπτωση η πρόταση της εναλλακτικής υπόθεσης απορρίπτεται (Κιντής, 2010). Ο έλεγχος των υποθέσεων αναλύεται διεξοδικά σε επόμενη ενότητα.

Στη συνέχεια είναι σημαντικός ο έλεγχος εγκυρότητας και σημαντικότητας του μοντέλου Πολλαπλής Γραμμικής Παλινδρόμησης χρησιμοποιώντας τον έλεγχο F. Προκειμένου να υπολογιστεί η τιμή F είναι αναγκαίο να κατασκευαστεί ο πίνακας ANOVA (Ανάλυση Διακυμάνσεων)[4]. Ο συγκεκριμένος έλεγχος αποτελεί έναν συνδυασμό των ελέγχων όλων των συντελεστών κλίσης των ανεξάρτητων

[4] Αναλύεται μέσα από την εφαρμογή στο SPSS σε επόμενη ενότητα.

μεταβλητών σε έναν. Δηλαδή πραγματοποιεί έλεγχο του ενδεχομένου ένας τουλάχιστον από τους συντελεστές κλίσης να είναι στατιστικά σημαντικά διαφορετικός από το μηδέν. Αξίζει να σημειωθεί ότι ο έλεγχος F είναι προτιμότερος από το να πραγματοποιήσουμε πολλούς ελέγχους t καθώς οι έλεγχοι t μπορεί να μας οδηγήσουν σε σφάλμα αν υποτεθεί γραμμικότητα ενώ στην πραγματικότητα ίσως να μην υπάρχει. Θα πρέπει να επιβεβαιωθεί η εγκυρότητα του μοντέλου μέσω του ελέγχου, αν δεν συμβεί αυτό θα πρέπει να απορρίψουμε το μοντέλο ή να το βελτιώσουμε. Επιπλέον, οι έλεγχοι t μπορεί να μη φανερώσουν τη γραμμική σχέση μεταξύ 2 ανεξάρτητων μεταβλητών δηλαδή να μη μπορεί να εντοπιστεί εξαρχής το πρόβλημα της πολυσυγγραμικότητας. Ενώ ο έλεγχος F μπορεί να εντοπίσει την ύπαρξη γραμμικής σχέσης και διορθώνει ως προς την πολυσυγγραμικότητα (Χάλκος, 2011). Παρακάτω παρουσιάζεται ο πίνακας ANOVA για μια πολλαπλή γραμμική παλινδρόμηση.

Πίνακας 4.1: ANOVA

	Βαθμοί Ελευθερίας	Άθροισμα Τετραγώνων	Μέσα Τετράγωνα	Στοιχεία Ελέγχου
Παλινδρόμηση	1	SSR	MSR	F=MSR/MSE
Κατάλοιπα	n-2	SSE	MSE	
Σύνολο	n-1	Μεταβλητότητα y		

Πηγή: (Χάλκος, 2011)

4.3 Συντελεστής πολλαπλής συσχέτισης

Οι συντελεστές συσχέτισης θα πρέπει να υπολογίζονται για όλα τα ζεύγη των μεταβλητών που περιλαμβάνουν την εξαρτημένη και τις ανεξάρτητες μεταβλητές. Στη περίπτωση που έχουμε μεταβλητές κανονικά κατανεμημένες, ο συντελεστής συσχέτισης υπολογίζεται με το τύπο του Pearson[5]. Αντίθετα στη περίπτωση που έχουμε πρόβλημα μη-κανονικότητας στις μεταβλητές που εξετάζουμε θεωρείται πιο κατάλληλος ο μη παραμετρικός συντελεστής συσχέτισης του Spearman με βάση του οποίου εξετάζουμε την ύπαρξη ισχυρής γραμμικής συσχέτισης είτε αυτή είναι θετική είτε αρνητική, ανάμεσα στην εξαρτημένη και στις ανεξάρτητες μεταβλητές. Ο συντελεστής του Spearman αποτελεί ενδεδειγμένη στατιστική όταν μια τουλάχιστον από τις μεταβλητές X και Y είναι μεταβλητή διάταξης. Ο συντελεστής Spearman είναι ανεξάρτητος από την κατανομή των X και Y. Μπορεί να χρησιμοποιηθεί και σε ποιοτικές μεταβλητές αλλά πρέπει να τις διατάξουμε σε κατηγορίες (Χάλκος, 2011).

Στην περίπτωση της απλής παλινδρόμησης (δηλαδή της παλινδρόμησης με μία επεξηγηματική μεταβλητή) παράλληλα με το συντελεστή προσδιορισμού R^2 ορίσαμε και τον συντελεστή γραμμικής συσχετίσεως (correlation coefficient) ρ και διαπιστώσαμε ότι ο συντελεστής προσδιορισμού ισούται με το τετράγωνο του συντελεστή συσχετίσεως. Στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης κατ' αναλογία με τον ρ ορίζεται ο συντελεστής πολλαπλής συσχέτισης R που είναι ένα μέτρο της συσχέτισης μεταξύ της Y και όλων των επεξηγηματικών μεταβλητών από κοινού. Δεδομένου ότι ο R ορίζεται ως η τετραγωνική ρίζα του αντίστοιχου

[5] Αναλύθηκε στην απλή γραμμική παλινδρόμηση

συντελεστή προσδιορισμού R^2 λαμβάνει μόνο θετικές τιμές, ή μηδέν, σε αντίθεση με τον ρ που όπως αναλύσαμε παραπάνω μπορεί να είναι θετικός, αρνητικός ή μηδέν. Αξίζει να τονιστεί ότι για την περίπτωση της πολλαπλής παλινδρόμησης αυτός που ενδιαφέρει είναι πρωτίστως ο R^2 και έπειτα ο ρ. Για τον συντελεστή προσδιορισμού θα μιλήσουμε αναλυτικότερα σε επόμενη ενότητα (Χάλκος, 2011).

4.4 Συντελεστής μερικής συσχέτισης

Ο συντελεστής μερικής συσχέτισης χρησιμοποιείται στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης όταν έχουμε παραπάνω από μια επεξηγηματικές μεταβλητές. Είναι εύλογο να ορίζουμε τους μερικούς συντελεστές συσχέτισης αντίστοιχα με τους μερικούς συντελεστές παλινδρόμησης.

Έτσι αν υποθέσουμε ότι έχουμε μια εξαρτημένη μεταβλητή Y και δυο ανεξάρτητες μεταβλητές X2 ,X3 τότε οι μερικοί συντελεστές συσχέτισης είναι οι παρακάτω:

- $r_{12,3}$= μερικός συντελεστής συσχετίσεως μεταξύ Y και X_2 κρατώντας τη X_3 σταθερή.
- $r_{13,2}$= μερικός συντελεστής συσχετίσεως μεταξύ Y και X_3 κρατώντας τη X_2 σταθερή.
- $r_{32,1}$= μερικός συντελεστής συσχετίσεως μεταξύ X_2 και X_3 κρατώντας τη Y σταθερή

Ο μερικός συντελεστής συσχέτισης μεταξύ Y και X_2 υπολογίζεται από την εξής σχέση:

$$r_{12,3} = \frac{\sum\{(\varepsilon_1 - \bar{\varepsilon_1}) \cdot (\varepsilon_2 - \bar{\varepsilon_2})\}}{\sqrt{\sum\{(\varepsilon_1 - \bar{\varepsilon_1}) \cdot (\varepsilon_2 - \bar{\varepsilon_2})\}}} = \frac{\sum(\varepsilon_1 \cdot \varepsilon_2)}{\sqrt{\sum(\varepsilon_1^2 \cdot \varepsilon_2)^2}}$$

(4.5)

όπου $\bar{\varepsilon_1}$ και $\bar{\varepsilon_2}$ αναμενόμενες τιμές της εξαρτημένης μεταβλητής 1 και ανεξάρτητης μεταβλητής 2. Σημειώνεται ότι ο δείκτης 1 αναφέρεται στη Y ο δείκτης 2 στη X_2 και ο δείκτης 3 στη X_3. Επιπρόσθετα οι μερικοί συντελεστές συσχέτισης μπορεί να εκφραστούν και άλλου τρόπου ως προς τους απλούς συντελεστές. Οι μερικοί συντελεστές συσχέτισης λέγονται και συντελεστές πρώτου βαθμού ο οποίος προσδιορίζει τον αριθμό των μεταβλητών που παραμένουν σταθερές όταν υπολογίζεται ο συντελεστής συσχετίσεως (Κιντής, 2010).

Για παράδειγμα αν θέλουμε να διερευνήσουμε την επίδραση που ασκούν η βροχή (X_2) και η θερμοκρασία (X_3) στην απόδοση μιας καλλιέργειας (Y) σε πρώτο στάδιο συμπεραίνουμε ότι η βροχή δεν σχετίζεται θετικά με την απόδοση της καλλιέργειας δηλαδή ισχύει r_{12}=0. Επιπλέον βλέπουμε ότι $r_{13} > 0$, $r_{23} < 0$. Από τη σχέση 4.5 αν τοποθετήσουμε σωστά τις μεταβλητές στην εξίσωση έχουμε ότι $r_{12.3} > 0$, δηλαδή διαπιστώνεται μία θετική συσχέτιση μεταξύ απόδοσης και βροχής. Το γεγονός αυτό οφείλεται στο ότι η τρίτη μεταβλητή δηλαδή η θερμοκρασία επιδρά πάνω στην απόδοση της καλλιέργειας αλλά και στη βροχόπτωση. Κατά συνέπεια για να αποδώσουμε την αληθή συσχέτιση μεταξύ της απόδοσης της καλλιέργειας και της βροχόπτωσης είναι αναγκαίο να απομονώσουμε την επίδραση της θερμοκρασίας. Οδηγούμαστε στο συμπέρασμα ότι οι μηδενικοί συντελεστές γραμμικής συσχέτισης μπορεί να εξάγουν εσφαλμένα συμπεράσματα (Montgomery, et al., 2012).

Το ποσοστό της μεταβλητότητας της Y που ερμηνεύεται από το υπόδειγμα παλινδρόμησης (δηλαδή από κοινού από τις X_2 και X_3) αποτελείται από τα εξής δύο μέρη: (α) το μέρος που ερμηνεύεται μόνο από τη X_2 (δηλαδή την r^2_{12}), και (β) το μέρος που δεν ερμηνεύεται από τη X_2 (δηλαδή 1- r^2_{12}) επί το ποσοστό που ερμηνεύεται από την X_3 κρατώντας σταθερή τη X_2.

Οι συντελεστές μερικής συσχέτισης έχουν μεγάλη σημασία για τον εντοπισμό των ψευδών συσχετίσεων. Η διαφορά μεταξύ του μερικού συντελεστή συσχέτισης και του απλού αποκτά σημασία για την εύρεση των λεγόμενων ψευδών συσχετίσεων (spurious correlations). Σύμφωνα με τη βιβλιογραφία υπάρχουν αρκετά άρθρα στα οποία χρησιμοποιούνται βασικές οικονομικές μεταβλητές όπως είναι το εισόδημα και ο πληθυσμός. Αυτά φαίνεται να συσχετίζονται σε πολλά παραδείγματα με μετεωρολογικές ή αστροφυσικές μεταβλητές. Σύμφωνα με τον David Hendry (1980) ο (μηδενικός) συντελεστής γραμμικής συσχετίσεως r_{IR} μεταξύ πληθωρισμού και αθροιστικής βροχόπτωσης στο H.B. είναι 0.98(!) (Hendry, 1980).

Για να αντιμετωπιστεί το πρόβλημα αυτό θα πρέπει να θεωρήσουμε ως επεξηγηματική μεταβλητή το χρόνο τόσο για τον πληθωρισμό όσο και για την αθροιστική βροχόπτωση. Κατά συνέπεια μπορούμε να εξαλείψουμε τη χρονική τάση που υποπτευόμαστε ότι ευθύνεται για την υψηλή (αλλά ψευδή) συσχέτιση μεταξύ των δύο μεταβλητών κάνοντας μια παλινδρόμηση για κάθε μια από τις μεταβλητές ξεχωριστά έχοντας ως ανεξάρτητη μεταβλητή το χρόνο. Έτσι έχουμε:

$$I = \hat{\beta}_1 + \hat{\beta}_2 t + \hat{u}_I \quad (4.6)$$

$$R = \hat{\beta}'_1 + \hat{\beta}'_2 t + \hat{u}_R \quad (4.7)$$

Όπου έχουμε I το πληθωρισμό, R την αθροιστική βροχόπτωση και t το χρόνο. Τα κατάλοιπα των δύο αυτών παλινδρομήσεων $\hat{u}_I$ και $\hat{u}_R$ δεν έχουν χρονική τάση και έτσι η μεταξύ τους συσχέτιση θα αποτυπώνει τη συσχέτιση των δύο αρχικών μεταβλητών που δεν οφείλεται στην χρονική τάση.

Κατά συνέπεια εκφράζει το μερικό συντελεστή συσχετίσεως (συντελεστή συσχετίσεως πρώτου βαθμού) μεταξύ πληθωρισμού και αθροιστικής βροχόπτωσης $\mathbf{r_{IR.t}}$. Συγκεκριμένα στο παράδειγμα είναι r_{IR}=0,98 αλλά $r_{IR.t}$ 0 (Hendry, 1980).

4.5 Συντελεστής πολλαπλού προσδιορισμού

Όπως και στην απλή παλινδρόμηση έτσι και στην πολλαπλή ο συντελεστής προσδιορισμού αποτελεί ένα κρίσιμο μέτρο και χρησιμοποιείται με σκοπό να αντιληφθούμε το βαθμό στον οποίο το μοντέλο μας ταιριάζει στα δεδομένα που έχουμε όσον αφορά την ερμηνεία του. Οι τιμές που παίρνει είναι από το 0 μέχρι το 1 και εκφράζεται και ως ποσοστό (Gerald, 2010).

Ο συντελεστής πολλαπλού προσδιορισμού χρησιμοποιείται για να μας δείξει το ποσό που αναλογεί από την εξηγήσιμη μεταβλητή του μοντέλου ως προς τη συνολική του μεταβλητότητα. Δηλαδή μας δείχνει σε τι ποσοστό η εξαρτημένη μεταβλητή του δείγματός μας ερμηνεύεται από τις ανεξάρτητες μεταβλητές. Όταν ο συντελεστής προσδιορισμού R^2 λαμβάνει υψηλή τιμή και μάλιστα κοντά στη μονάδα σημαίνει ότι υπάρχει μεγάλος βαθμός ερμηνείας της μεταβλητότητας της εξαρτημένης μεταβλητής από τις ανεξάρτητες μεταβλητές. Αυτό μας οδηγεί στο συμπέρασμα ότι γνωρίζοντας τις τιμές των ανεξάρτητων μεταβλητών του δείγματός

μας μπορούμε να προβλέψουμε αρκετά καλά τις τιμές που παίρνει η εξαρτημένη μεταβλητή (Κιντής, 2010).

Γίνεται φανερό ότι προσθέτοντας παραπάνω επεξηγηματικές μεταβλητές η τιμή του συντελεστή πολλαπλού προσδιορισμού θα αυξάνεται ή στην ακραία περίπτωση θα παραμείνει αμετάβλητη. Αυτό αποτελεί ίσως ένα από τα μειονεκτήματα του συντελεστή προσδιορισμού.

Τη λύση σε αυτό το πρόβλημα έδωσε αρχικά ο H. Theil ο οποίος πρότεινε να χρησιμοποιηθούν διακυμάνσεις αντί μεταβλητότητες στον ορισμό του R^2. Στο σημείο αυτό αξίζει να θυμηθούμε ότι οι διακυμάνσεις εκφράζουν μεταβλητότητα κατά βαθμό ελευθερίας. Συνεπώς ορίζεται ο λεγόμενος διορθωμένος συντελεστής προσδιορισμού (adjusted coefficient of determination) και υπολογίζεται ως εξής:

$$\tilde{R}^2 = \frac{1 - \textit{διακύμανση καταλοίπων}}{\textit{διακύμανση της } Y} \quad \textbf{(4.8)}$$

Επίσης ένας μελετητής θα πρέπει να προσέξει αν θέλει να συγκρίνει δύο ή περισσότερα υποδείγματα παλινδρόμησης με βάση την τιμή του συντελεστή προσδιορισμού τους (διορθωμένου ή μη) να είναι τα ίδια η εξαρτημένη μεταβλητή με το μέγεθος του δείγματος. Οι επεξηγηματικές μεταβλητές θα πρέπει πάντα να είναι μικρότερες σε αριθμό από το μέγεθος του δείγματος (αριθμός παρατηρήσεων) και να υπεισέρχονται στο υπόδειγμα με οποιαδήποτε μορφή (Χάλκος, 2011).

4.6 Πολλαπλή γραμμική παλινδρόμηση με χρήση του SPSS

Το πρώτο βήμα για να πραγματοποιήσουμε μια πολλαπλή γραμμική παλινδρόμηση είναι να ορίσουμε το δείγμα μας και να καταχωρήσουμε τα δεδομένα που θα χρησιμοποιήσουμε στο SPSS. Σε πρώτο στάδιο διερευνώνται οι μεταβλητές που περιλαμβάνονται στο δείγμα μας και οι οποίες θα συμμετέχουν στην εκτέλεση της πολλαπλής γραμμικής παλινδρόμησης. Στον Πίνακα 4.2 παρουσιάζονται τα δεδομένα του παραδείγματος που θα χρησιμοποιήσουμε προκειμένου να εξάγουμε τα αποτελέσματα από την ΠΓΠ. Τα δεδομένα αφορούν σωματικές μετρήσεις 40 ατόμων δηλαδή το δείγμα μας αποτελείται από 20 παρατηρήσεις. Θέλουμε να προβλέψουμε το δείκτη σωματικού λίπους μέσα από τις σωματικές μετρήσεις. Η εξαρτημένη μεταβλητή μας είναι ο δείκτης σωματικού λίπους και οι ανεξάρτητες μεταβλητές είναι το πάχος του δέρματος, η περιφέρεια του μηρού και η περιφέρεια μπράτσου (Χατζηνικολάου, 2002).

Πίνακας 4.2: Δεδομένα της εφαρμογής της ΠΓΠ

πάχος δέρματος	περιφέρεια μηρού	περιφέρεια μπράτσου	δείκτης σωματικού λίπους
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

Οι μονάδες μέτρησης της εξαρτημένης μεταβλητής είναι ένας καθαρός αριθμός ενώ των ανεξάρτητων μεταβλητών είναι τα εκατοστά. Αξίζει να σημειωθεί εδώ ότι το βασικό πλεονέκτημα της παλινδρόμησης είναι ότι δίνει τη δυνατότητα του ελέγχου της στατιστικής σημαντικότητας μεταξύ μεταβλητών που μπορεί να μην έχουν την ίδια κλίμακα μέτρησης.

Για να πάρουμε τα περιγραφικά στατιστικά μέτρα των μεταβλητών και να τις κατανοήσουμε ακόμα καλύτερα ακολουθούμε την ίδια διαδικασία όπως παραθέσαμε και στην απλή παλινδρόμηση. Δηλαδή στο SPSS ακολουθούμε την εξής διαδικασία στη γραμμή εργαλείων Analyze-Descriptive Statistics και επιλέγουμε τις μεταβλητές που μας ενδιαφέρουν. Εδώ επιλέγουμε όλες τις μεταβλητές και θα πάρουμε τα εξής περιγραφικά στατιστικά μέτρα:

Πίνακας 4.3: Περιγραφικά στατιστικά μέτρα

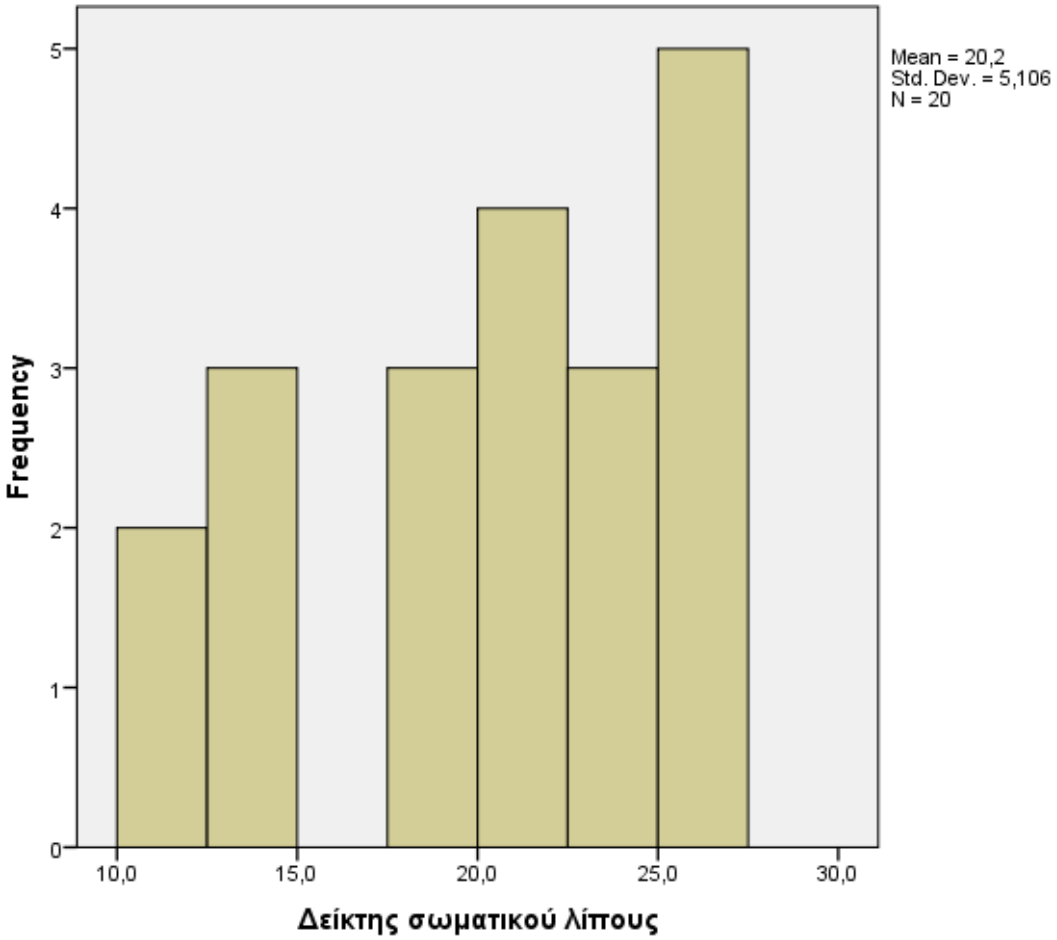
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Πάχος δέρματος	20	14,6	31,4	25,305	5,0233
Περιφέρεια μηρού	20	42,2	58,6	51,170	5,2346
Περιφέρεια μπράτσου	20	21,3	37,0	27,620	3,6471
Δείκτης σωματικού λίπους	20	11,7	27,2	20,195	5,1062
Valid N (listwise)	20				

Στον Πίνακα 4.3 παρουσιάζονται τα περιγραφικά στατιστικά μέτρα των δεδομένων που χρησιμοποιούμε. Βλέπουμε ότι ο αριθμός των παρατηρήσεων N είναι 20. Παρατηρούμε ότι δεν υπάρχουν τιμές που να λείπουν (missing values). Επίσης παίρνουμε τα στατιστικά μέτρα που αφορούν την ελάχιστη τιμή, τη μέγιστη, τη μέση τιμή και την τυπική απόκλιση των μεταβλητών που εξετάζουμε. Ο μέσος όρος δείχνει το επίπεδο των τιμών στο οποίο κυμαίνονται τα δεδομένα, ενώ η ελάχιστη, μέγιστη τιμή καθώς και η τυπική απόκλιση δείχνουν τον τρόπο που τα δεδομένα εξαπλώνονται γύρω από το μέσο όρο. Η περιφέρεια μηρού παρουσιάζει το μεγαλύτερο μέσο όρο 51,1 εκατοστά με μέγιστη τιμή τα 58,6 εκατοστά. Η ελάχιστη τιμή του δείκτη σωματικού λίπους είναι 11,7 κιλά λίπους.

Το επόμενο βήμα είναι να ελέγξουμε ως προς την κανονικότητα την εξαρτημένη μεταβλητή δηλαδή το δείκτη σωματικού λίπους. Τη κανονικότητα μπορούμε να την απεικονίσουμε με γραφικό τρόπο μέσα από ένα Ιστόγραμμα. Για να επιβεβαιωθεί η στατιστική σημαντικότητα εφαρμόζουμε έναν μη παραμετρικό έλεγχο ο οποίος ονομάζεται Kolmogorov-Smirnov όπως θα δούμε παρακάτω. Πατάμε Graphs→Legacy Dialogs→Histogram, βάζουμε στο κουτάκι την εξαρτημένη μεταβλητή που θέλουμε να εξετάσουμε τη κανονικότητά της δηλαδή στη προκειμένη περίπτωση το δείκτη σωματικού λίπους και έχουμε το εξής ιστόγραμμα:

Ιστόγραμμα 1: Μελέτη κανονικότητας εξαρτημένης μεταβλητής



Από το παραπάνω ιστόγραμμα αντιλαμβανόμαστε την κατανομή των δεδομένων που χρησιμοποιούμε. Ο μη παραμετρικός έλεγχος Kolmogorov-Smirnov που εφαρμόσαμε μας κατευθύνει στο αν δεν απορρίψουμε ή απορρίψουμε την αρχική μας υπόθεση ότι έχουμε μια κανονική κατανομή. Παρατηρώντας τη τιμή του sig. στον παρακάτω Πίνακα μπορούμε να απορρίψουμε ή όχι την αρχική μας υπόθεση. Στη περίπτωση που η τιμή sig. υπερβαίνει το 0,05 τότε η κατανομή είναι κανονική. Στο παράδειγμά μας διαπιστώνουμε ότι η Κατανομή του δείκτη σωματικού λίπους (lipos) δεν είναι κανονική.

Για να εξάγουμε έλεγχο κανονικότητας στο SPSS ακολουθούμε τα εξής βήματα: Analyze→Nonparametric Tests→1- Sample K-S. Περνάμε στο κουτάκι που μας βγάζει την εξαρτημένη μεταβλητή προκειμένου να γίνει ο έλεγχος του ενδεχομένου οι τιμές της να ακολουθούν την κανονική κατανομή. Να σημειωθεί ότι η επιλογή για τον έλεγχο κανονικότητας έχει ήδη προεπιλεχθεί από το SPSS (Normal). Αν επιλέξουμε το κουτάκι Options εμφανίζεται ένα άλλο παράθυρο στο οποίο μπορούμε να επιλέξουμε να εμφανιστεί ένας πίνακας που να περιλαμβάνει κάποια περιγραφικά μέτρα αυτών των μεταβλητών. Πατάμε οκ και εμφανίζεται ο παρακάτω πίνακας.

Πίνακας 4.4: One-Sample Kolmogorov-Smirnov Test

One-Sample Kolmogorov-Smirnov Test

		Δείκτης σωματικού λίπους
N		20
Normal Parameters[a,b]	Mean	20,195
	Std. Deviation	5,1062
Most Extreme Differences	Absolute	,123
	Positive	,123
	Negative	-,120
Kolmogorov-Smirnov Z		,552
Asymp. Sig. (2-tailed)		,921

a. Test distribution is Normal.

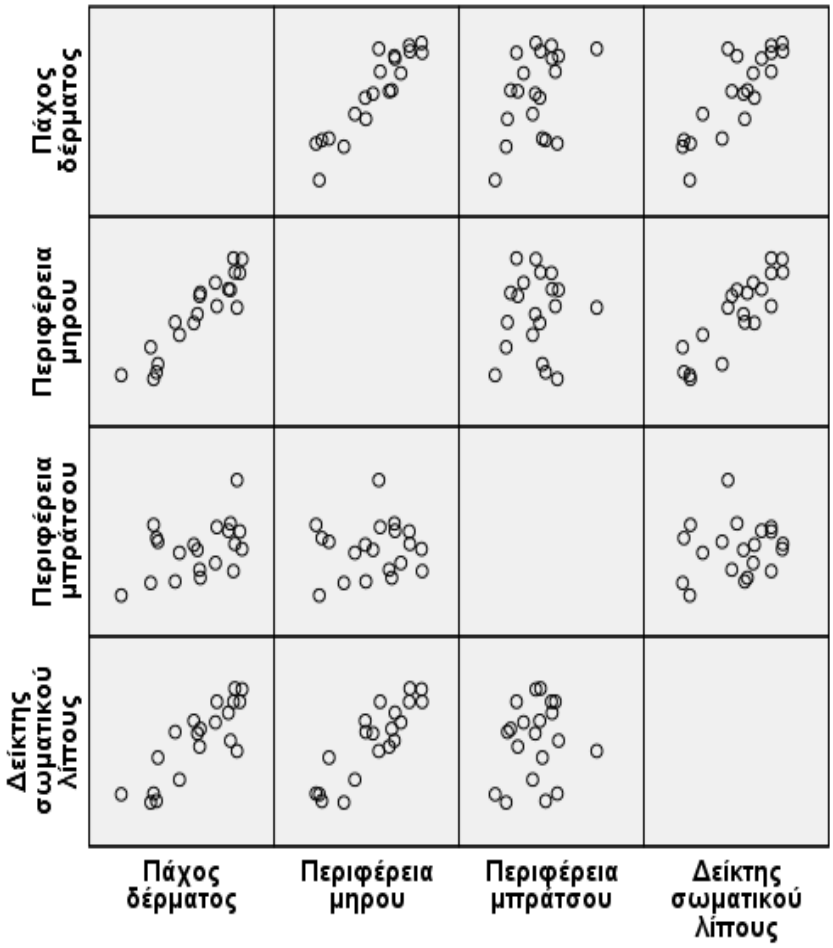
b. Calculated from data.

Διαπιστώνουμε ότι η κατανομή του δείκτη σωματικού λίπους είναι κανονική διότι sig=0,921>0,05. Ο έλεγχος της υπόθεσης της κανονικότητας είναι δυνατό να οδηγήσει ή στην αποδοχή ή στην απόρριψή της. Στην περίπτωση που αποδεχόμαστε την υπόθεση της κανονικότητας προχωράμε στον έλεγχο των επόμενων υποθέσεων. Στη περίπτωση που η υπόθεση μας δεν είναι αποδεκτή τότε οφείλουμε να τροποποιήσουμε τα δεδομένα έτσι ώστε να ακολουθούν κανονική κατανομή. Θυμίζουμε ότι ο έλεγχος της κανονικότητας είναι απαραίτητος ώστε να γνωρίζει ο ερευνητής αν η μέθοδος είναι έγκυρη ή όχι. Στο παράδειγμα που εξετάζουμε για αριθμό παρατηρήσεων N=20 τα δεδομένα της εξαρτημένης μεταβλητής τείνουν να ακολουθούν την κανονική κατανομή.

Το επόμενο σημαντικό βήμα για να προχωρήσει η διαδικασία της πολλαπλής γραμμικής παλινδρόμησης είναι η εξέταση της σχέσης σύνδεσης των μεταβλητών που συμμετέχουν στην ανάλυσή μας. Δηλαδή θέλουμε να εξετάσουμε αν οι ανεξάρτητες μεταβλητές που χρησιμοποιούνται για την επεξήγηση της συμπεριφοράς της εξαρτημένης μεταβλητής συνδέονται γραμμικά με το εξαρτημένο μέγεθος.

Αρχικά επιδιώκεται ο εντοπισμός της ύπαρξης μιας τέτοιας σχέσης μέσα από τη γραφική απεικόνιση των μεταβλητών και επιβεβαιώνεται στατιστικά μέσω των πινάκων συσχετίσεων όπως θα δούμε παρακάτω. Σε ένα πολλαπλό διάγραμμα σημείων (επιλέγουμε scatter plot στο SPSS και τοποθετούμε τις μεταβλητές ενδιαφέροντος) μπορούμε να παρατηρήσουμε ταυτόχρονα την γραφική απεικόνιση των σχέσεων των ανεξάρτητων μεταβλητών με την εξαρτημένη.

4.1 Διάγραμμα πολλαπλής απεικόνισης



Από το Διάγραμμα 4.1 διαπιστώνεται ότι σημειώνεται καλύτερη γραμμική σχέση μεταξύ των μεταβλητών του Πάχους δέρματος και της Περιφέρειας του μηρού. Ενώ παρατηρούμε ότι η σχέση μεταξύ της εξαρτημένης μεταβλητής του Δείκτη σωματικού λίπους και των ανεξάρτητων μεταβλητών θα ήταν προτιμότερο στη βέλτιστη περίπτωση να ήταν μια ευθεία γραμμή. Στη πράξη αυτό συμβαίνει πολύ δύσκολα. Το αξιοσημείωτο που πρέπει να διερευνηθεί είναι η ύπαρξη κάποιας άλλης κίνησης π.χ. περιοδικής.

Η εφαρμογή της μεθόδου της γραμμικής παλινδρόμησης αποσκοπεί στο να περιγράψουμε την κίνηση της εξαρτημένης μεταβλητής με βάση τις τιμές κάποιων επεξηγηματικών (ανεξάρτητων) μεταβλητών. Για το λόγο αυτό θέλουμε η κάθε ανεξάρτητη μεταβλητή να συσχετίζεται σε ισχυρό βαθμό με την εξαρτημένη. Όταν υπάρχει υψηλός βαθμός συσχέτισης οδηγούμαστε στο γεγονός ότι οι περισσότερες πληροφορίες της εξαρτημένης μεταβλητής εξηγούνται από την ανεξάρτητη μεταβλητή του δείγματος. Όμως στη περίπτωση που έχουμε μεγάλο αριθμό ανεξάρτητων μεταβλητών καλό είναι οι ανεξάρτητες μεταβλητές να μην συσχετίζονται ισχυρά μεταξύ τους. Όταν οι ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους τότε δίνουν την ερμηνεία του ίδιου μέρους της διακύμανσης της εξαρτημένης μεταβλητής. Όταν συμμετέχουν παραπάνω από δυο ανεξάρτητες συσχετισμένες μεταβλητές οδηγούμαστε στην αύξηση της πιθανότητας λάθους στο μοντέλο που εξετάζουμε. Για αυτό το λόγο είναι εύλογο πριν εφαρμόσουμε τη γραμμική παλινδρόμηση να εξετάζουμε τις συσχετίσεις που δημιουργούνται μεταξύ

των μεταβλητών. Ο Πίνακας συσχετίσεων περιλαμβάνει την εξαγωγή των αποτελεσμάτων κατά τη μελέτη των συσχετίσεων. Σε αυτού του είδους τους πίνακες παρατηρούμε την τιμή του δείκτη η οποία περιγράφει την ύπαρξη της συσχέτισης καθώς και τη τιμή του ελέγχου ως προς τη σημαντικότητα της τιμής του συγκεκριμένου συντελεστή (Κιντής, 2010).

Για να εξάγουμε στο SPSS τους πίνακες συσχετίσεων ακολουθούμε τα βήματα Analyze→Correlate→Bivariate. Θα υπολογίσουμε τους συντελεστές γραμμικής συσχέτισης για όλα τα ζεύγη των μεταβλητών. Επιλέγουμε στα κουτιά τις μεταβλητές ενδιαφέροντος συγκεκριμένα εδώ επιλέγουμε όλες τις μεταβλητές που εξετάζουμε και στη συνέχεια στις επιλογές που μας δίνει το πρόγραμμα για τους συντελεστές συσχέτισης που θέλουμε να μας δώσουν τις συσχετίσεις. Εμείς επιλέγουμε το συντελεστή Pearson και στο από κάτω κουτάκι για το test of significance επιλέγουμε Two-tailed. Αν θέλουμε να εμφανιστούν και οι άλλοι δύο συντελεστές απλά τους επιλέγουμε. Παρατηρούμε ότι κάτω αριστερά είναι επιλεγμένη μία επιλογή (Flag significant correlations). Η επιλογή Options μας δίνει τη δυνατότητα εμφάνισης των μέσων, των τυπικών αποκλίσεων και των πληθών των τιμών για κάθε μεταβλητή. Για να πάρουμε τον πίνακα συσχετίσεων πατάμε οκ και έχουμε τον εξής πίνακα:

Πίνακας 4.5: Συσχετίσεις μεταβλητών ανά ζεύγη

Correlations

		Πάχος δέρματος	Περιφέρεια μηρου	Περιφέρεια μπράτσου	Δείκτης σωματικού λίπους
Πάχος δέρματος	Pearson Correlation	1	,924**	,458*	,843**
	Sig. (2-tailed)		,000	,042	,000
	N	20	20	20	20
Περιφέρεια μηρου	Pearson Correlation	,924**	1	,085	,878**
	Sig. (2-tailed)	,000		,723	,000
	N	20	20	20	20
Περιφέρεια μπράτσου	Pearson Correlation	,458*	,085	1	,142
	Sig. (2-tailed)	,042	,723		,549
	N	20	20	20	20
Δείκτης σωματικού λίπους	Pearson Correlation	,843**	,878**	,142	1
	Sig. (2-tailed)	,000	,000	,549	
	N	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

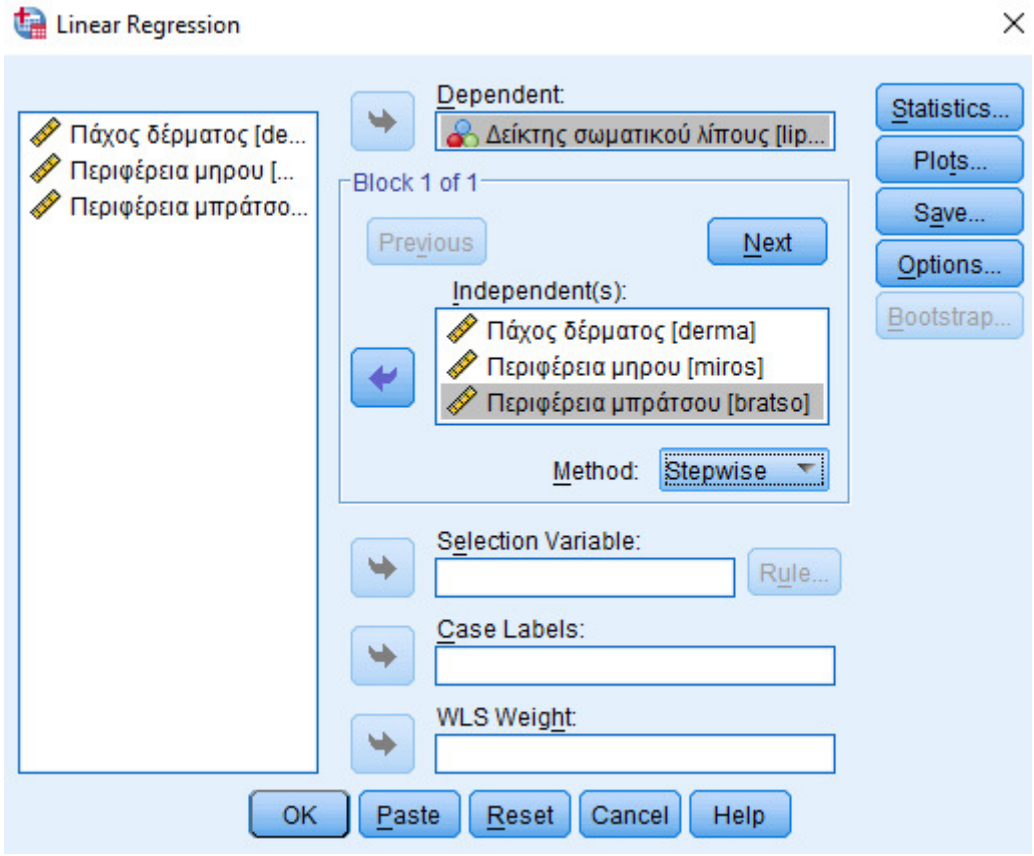
*. Correlation is significant at the 0.05 level (2-tailed).

Στον Πίνακα 4.5 παρουσιάζονται οι συσχετίσεις των εξεταζόμενων μεταβλητών ανά ζεύγη. Παρατηρούμε για N=20 ότι από τις ανεξάρτητες μεταβλητές η μεταβλητή Περιφέρεια μηρού συσχετίζεται πιο έντονα με την εξαρτημένη μεταβλητή το δείκτη σωματικού λίπους. Επίσης διαπιστώνουμε ότι οι ανεξάρτητες μεταβλητές περιφέρεια μπράτσου και πάχος δέρματος καθώς και η περιφέρεια μηρού με το πάχος δέρματος είναι πολύ ισχυρά συσχετισμένες μεταξύ τους. Μετά από τον έλεγχο των συσχετίσεων θα αποφασίσουμε αν πρόκειται να χρησιμοποιήσουμε όλες τις ανεξάρτητες μεταβλητές για την κατασκευή του μοντέλου ή αν θα επιλέξουμε κάποιες από αυτές.

Έτσι είμαστε έτοιμοι να προχωρήσουμε στην κατασκευή του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης. Έχουμε δηλαδή τη δυνατότητα να εντοπίσουμε τις στατιστικά σημαντικές μεταβλητές για το εξαρτημένο μέγεθος καθώς και να προσδιορίσουμε το μοντέλο το οποίο περιγράφει τον τρόπο που συνδέονται οι μεταβλητές μεταξύ τους.

Ακολουθώντας τα εξής βήματα Analyze → Regression→ Linear Regression όπως και στην απλή γραμμική παλινδρόμηση ενεργοποιούμε το πλαίσιο διαλόγου μέσα στο οποίο θα ορίσουμε τις μεταβλητές που θα συμμετέχουν στην πολλαπλή γραμμική παλινδρόμηση. Στο εξεταζόμενο παράδειγμα ορίζουμε ως εξαρτημένη μεταβλητή το δείκτη σωματικού λίπους και σαν ανεξάρτητες όλες τις μεταβλητές που έχουμε ήδη αναφέρει. Υπάρχουν διάφοροι τρόποι κατασκευής του μοντέλου της παλινδρόμησης. Η κάθε διαδικασία κατασκευής του μοντέλου ακολουθεί την δική της λογική και σχεδόν ποτέ δεν καταλήγουν στον ίδιο αποτέλεσμα (Τσαγρής, 2008).

Εικόνα 4.1: Κατασκευή του μοντέλου της ΠΓΠ



Στο κουτάκι method θα πρέπει να επιλέξουμε τις μεθόδους που θα χρησιμοποιήσουμε για την εξαγωγή της ΠΓΠ. Οι συχνότερα χρησιμοποιούμενες μέθοδοι είναι οι εξής: Enter, Forward, Backward και Stepwise. Με τη μέθοδο Enter χρησιμοποιούμε στο μοντέλο μας όλες τις μεταβλητές οι οποίες προτείνονται από τον αναλυτή. Προκειμένου να χρησιμοποιηθεί αυτή η τεχνική ο χρήστης είναι αναγκαίο να προχωρήσει στην απόκλιση από την ανάλυση κάποιων ανεξάρτητων μεταβλητών που έχουν ισχυρή σύνδεση μεταξύ τους.

Με τη μέθοδο Forward εντοπίζουμε ανάμεσα στις ανεξάρτητες μεταβλητές εκείνη τη μεταβλητή που παρουσιάζει ισχυρή σύνδεση με την εξαρτημένη

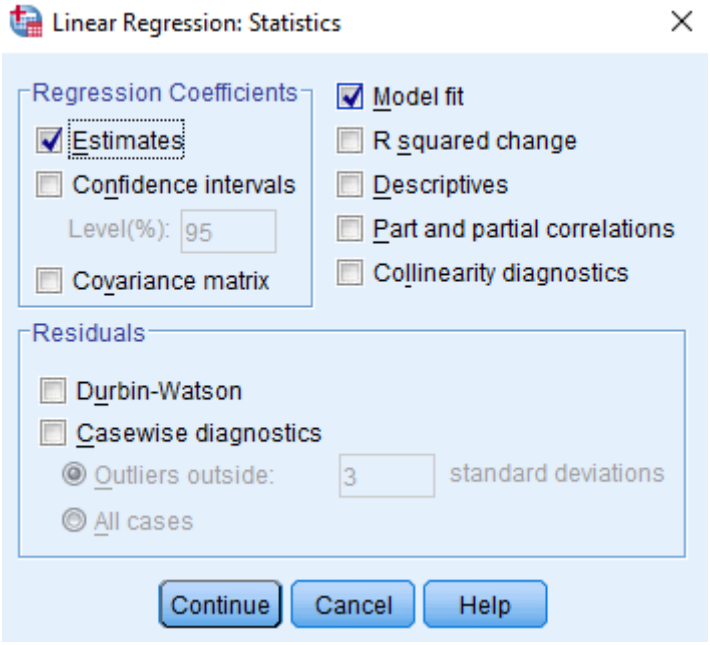
μεταβλητή. Είναι υπεύθυνη για τον στατιστικό έλεγχο ως προς τη σημαντικότητα της συγκεκριμένης μεταβλητής καθώς και έναν έλεγχο για τη συνολική εικόνα του μοντέλου. Επιπλέον σε δεύτερο στάδιο προχωράει στην επιλογή της μεταβλητής αυτής που παρουσιάζει την αμέσως υψηλότερη συσχέτιση με την εξαρτημένη καθώς και την αμέσως χαμηλότερη συσχέτιση με την ανεξάρτητη που υπάρχει ήδη στο μοντέλο. Στη συνέχεια γίνονται οι έλεγχοι που σχετίζονται με την στατιστική σημαντικότητα των μεταβλητών και ελέγχεται το μοντέλο ως προς τη σημαντικότητά του. Η παραπάνω διαδικασία γίνεται αρκετές φορές σε επανάληψη έως ότου να μην υφίσταται άλλη στατιστικά σημαντική μεταβλητή ως προς την εξαρτημένη (Τσαγρής, 2008).

Η μέθοδος Backward είναι η αντίθετη από την μέθοδο Forward. Στο πρώτο στάδιο εισάγουμε όλες τις μεταβλητές στο μοντέλο και σε κάθε βήμα που ακολουθεί πραγματοποιείται αφαίρεση της μεταβλητής που θεωρείται λιγότερο στατιστικά σημαντική για την εξαρτημένη μεταβλητή του μοντέλου μας. Η διαδικασία γίνεται πολλές φορές έως ότου καταλήξουμε μόνο σε στατιστικά σημαντικές μεταβλητές που θέλουμε να συμμετέχουν στο μοντέλο μας. Επίσης είναι σημαντικό αυτές οι μεταβλητές να μη συσχετίζονται μεταξύ τους αλλά να είναι ισχυρά συσχετισμένες με την εξαρτημένη μεταβλητή (Montgomery, et al., 2012).

Η μέθοδος Stepwise αποτελεί τη πιο γνωστή μέθοδο και συνδυάζει τη μέθοδο Forward και τη μέθοδο Backward. Μέσα από τους διαφορετικούς τρόπους κατασκευής των μοντέλων οδηγούμαστε και σε διαφορετικά μοντέλα. Το γεγονός αυτό συμβαίνει διότι η παλινδρόμηση αποτελεί μια πολυπαραγοντική μέθοδο στην οποία μια μεταβλητή που προστίθεται ή αφαιρείται είναι ικανή να αυξήσει ή να μειώσει η τις τιμές των συντελεστών συσχέτισης. Υπάρχει η δυνατότητα να προσδιορίσουμε έναν αριθμό από δείκτες οι οποίοι βοηθούν στη καλή προσαρμογή του μοντέλου στα δεδομένα μας. Επίσης μέσα από τα γραφήματα και τους πίνακες που συνοδεύουν τους παραπάνω δείκτες μπορούμε να κατανοήσουμε καλύτερα τα δεδομένα μας και κατά συνέπεια το μοντέλο μας (Ιωαννίδης, 2005).

Σκοπός της κατασκευής του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης είναι ο έλεγχος της στατιστικής σημαντικότητας αυτού, ο έλεγχος της προσαρμογής των εξεταζόμενων δεδομένων στο εξαρτημένο μέγεθος και η επίδραση που ασκούν οι ανεξάρτητες μεταβλητές πάνω στο εξαρτημένο μέγεθος. Συνεπώς στην Εικόνα 4.1 επιλέγουμε το πλήκτρο Statistics και ενεργοποιούμε το πλαίσιο διαλόγου που παρουσιάζεται στην Εικόνα 4.2 και από το οποίο ορίζουμε συμπληρωματικούς στατιστικούς δείκτες.

Εικόνα 4.2 : Statistics - ΠΓΠ

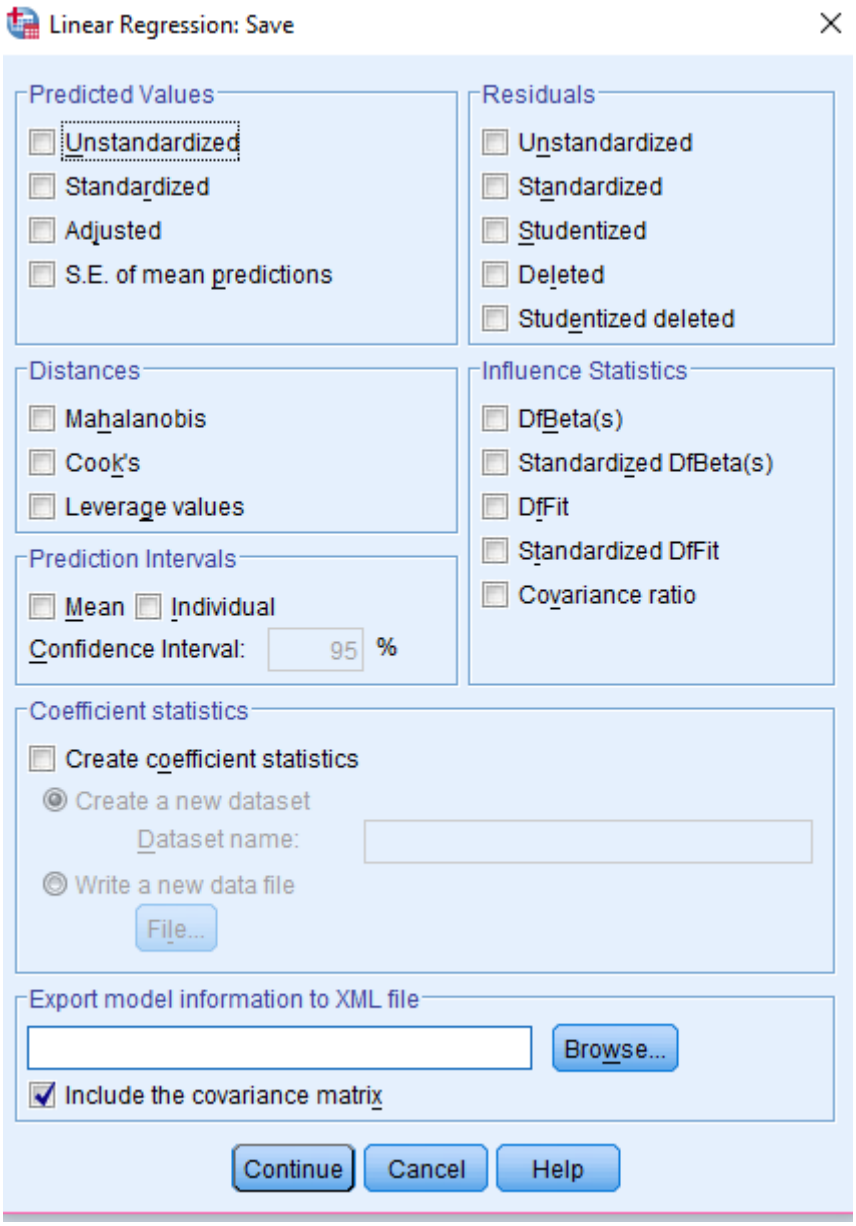


Στην Εικόνα 4.2 βλέπουμε τα μέτρα που σχετίζονται με τους «συντελεστές της παλινδρόμησης» και είναι τα εξής:

- Estimates: είναι οι συντελεστές που χρησιμοποιούνται στο μοντέλο καθώς και στατιστικά μέτρα που σχετίζονται με αυτούς τους συντελεστές. Αυτά τα μέτρα είναι η διακύμανση, το τυπικό σφάλμα κλπ.
- Confidence Interval: είναι το διάστημα εμπιστοσύνης στο 95% που χρησιμοποιείται για να εκτιμηθεί η τιμή των συντελεστών της παλινδρόμησης.
- Covariance Matrix: αποτελεί το πίνακα συνδιακύμανσης και συσχέτισης.
- Model fit: από τη συγκεκριμένη επιλογή ορίζουμε και τον πίνακα ANOVA ο οποίος περιλαμβάνει τον κύριο έλεγχο που αφορά την στατιστική σημαντικότητα όλου του μοντέλου.
- R squared change: είναι ο συντελεστής προσδιορισμού. Η συγκεκριμένη επιλογή μας παρέχει πληροφορίες που σχετίζονται με την αλλαγή της τιμής του P2 προσθέτοντας ή διαγράφοντας μια ανεξάρτητη μεταβλητή.
- Desriptives: αυτή η επιλογή μας βοηθάει να πάρουμε αποτελέσματα που αφορούν τα περιγραφικά στατιστικά μέτρα των δεδομένων μας
- Part and Partial correlation: δείχνει το δείκτη αυτοσυσχέτισης και μερικής αυτοσυσχέτισης των εξεταζόμενων μεταβλητών
- Collinearity diagnostics: μελετάται η συγραμμικότητα μέσα από την εξαγωγή ενός μεγάλου αριθμού στατιστικών μέτρων. Εδώ αναφερόμαστε στο γεγονός όπου η εξαρτημένη μεταβλητή της ανάλυσής μας είναι γραμμικός συνδυασμός μιας άλλης εξαρτημένης μεταβλητής
- Durbin- Watson : αποτελεί ένα σημαντικό στατιστικό έλεγχο ο οποίος ελέγχει ως προς την σειριακή συσχέτιση μεταξύ των καταλοίπων. Με την εφαρμογή αυτού του ελέγχου παίρνουμε τα περιγραφικά στατιστικά μέτρα τόσο των καταλοίπων όσο και των προβλεπόμενων τιμών από το μοντέλο
- Casewise diagnostics: περιγράφει τη συμπεριφορά που εμφανίζουν οι ακραίες τιμές (Εικόνα 4.3).

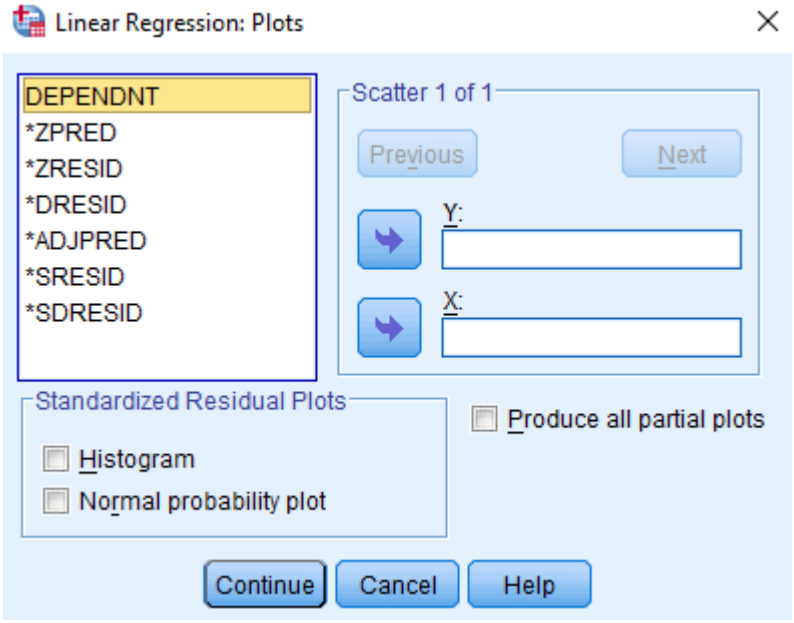
Πατώντας το πλήκτρο Save στην Εικόνα 4.1 εμφανίζεται το πλαίσιο στην Εικόνα 4.3.

Εικόνα 4.3: Save στο SPSS



Στη συνέχεια πατώντας το πλήκτρο plot όπως βλέπουμε στο κουτάκι στην Εικόνα 4.1 ενεργοποιείται το παράθυρο το οποίο επιτρέπει την δημιουργία γραφημάτων. Μέσα από αυτό το παράθυρο επιλογών μπορούμε να δημιουργήσουμε διάφορα γραφήματα όπως είναι το ιστόγραμμα για τα κανονικοποιημένα κατάλοιπα. Το παράθυρο που εμφανίζεται στο SPSS απεικονίζεται στην Εικόνα 4.4.

Εικόνα 4.4 Επιλογή Normal probability plot



Στη παραπάνω Εικόνα επιλέγουμε τα κανονικοποιημένα κατάλοιπα για να πάρουμε το διάγραμμα των καταλοίπων το οποίο χρησιμοποιείται για να ελεγχθούν ως προς την κανονικότητα τα κατάλοιπα. Στην περίπτωση που επιλέξουμε το διπλανό κουτάκι δηλαδή την επιλογή Produce all partial plots θα παράγουμε τα scatterplots για κάθε μια από τις ανεξάρτητες μεταβλητές με τα κατάλοιπα της εξαρτημένης. Στις επιλογές X και Y μπορούμε να επιλέξουμε από την λίστα αριστερά, όποιο ζευγάρι μεταβλητών θέλουμε προκειμένου να κατασκευάσουμε τα scatter plots (Κιντής, 2010).

Στην Εικόνα 4.1 πατώντας το πλήκτρο Save επιλέγουμε εκείνα τα μεγέθη τα οποία επιθυμούμε να αποθηκεύσουμε στο αρχείο των δεδομένων μας με τη μορφή μεταβλητών. Οι επιλογές που έχουμε παρουσιάζονται παρακάτω.

Πίνακας 4.6 : Αποθήκευση μεταβλητών στο αρχείο δεδομένων

Keyword	Statistic
dependnt	dependent variable
*zpred	standardized predicted values
*zresid	standardized residuals
*dresid	deleted residuals
*adjpred .	adjusted predicted values
*sresid	studentized residuals
*sdresid	studentized deleted residuals

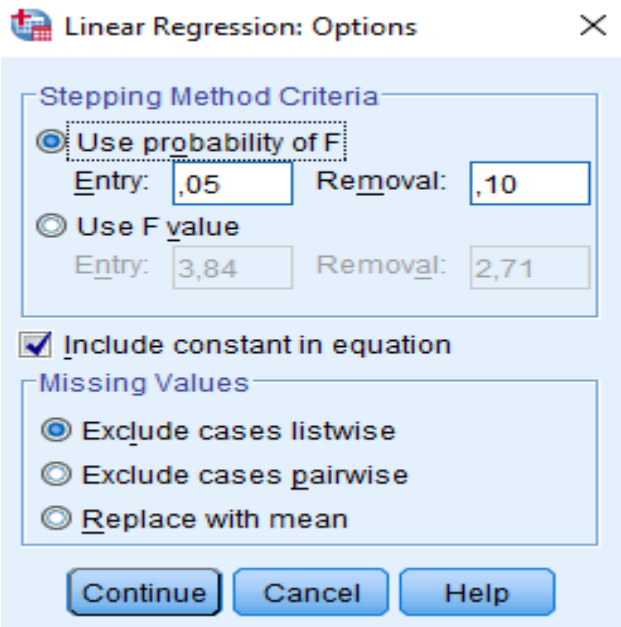
Από τον παραπάνω πίνακα αξίζει να σημειωθούν τα εξής:

- Unstantarized Pridicted Values: η επιλογή αυτή μας βοηθά να αποθηκεύσουμε τις τιμές της εξαρτημένης μεταβλητής έτσι όπως την υπολογίζουμε από την εξίσωση παλινδρόμησης.
- Stantarized Pridicted Values: με αυτή την επιλογή αποθηκεύουμε τις κανονικοποιημένες προβλεπόμενες τιμές.

- Adjusted: είναι οι προσαρμοσμένες τιμές δηλαδή αποθηκεύονται οι προβλεπόμενες τιμές για κάθε εγγραφή που εξαιρείται κατόπιν του υπολογισμού των συντελεστών παλινδρόμησης.
- Unstantarized Residual: αποθηκεύονται τα κατάλοιπα
- Stantarized Residual: αποθηκεύονται οι κανονικοποιημένες τιμές των καταλοίπων
- Mahalanobis: είναι ένα μέτρο που μετρά τη διαφορά μεταξύ της εγγραφής της εξαρτημένης μεταβλητής και του μέσου όρου των συνολικών εγγραφών.
- Cook's: αποθηκεύεται το μέτρο που δείχνει την αλλαγή των καταλοίπων μιας εγγραφής στην περίπτωση που αυτή η εγγραφή αποκλειστεί από τον υπολογισμό των συντελεστών παλινδρόμησης.

Τέλος πριν προχωρήσουμε στην ερμηνεία των αποτελεσμάτων αναφέρουμε ότι με την επιλογή Options στην Εικόνα 4.1 ορίζουμε το επίπεδο σημαντικότητας για την απόρριψη ή όχι του μοντέλου παλινδρόμησης. Συνήθως το επίπεδο στατιστικής σημαντικότητας που χρησιμοποιείται είναι το α=0.05.

Εικόνα 4.5: Επιλογή στατιστικής σημαντικότητας



Επίσης όπως βλέπουμε στην παραπάνω Εικόνα μας δίνεται η δυνατότητα να αφαιρέσουμε το σταθερό όρο από την εξίσωση παλινδρόμησης. Με την επιλογή Missing Values μπορούμε να εξετάσουμε και να επέμβουμε στον τρόπο χρήσης των ελλειπουσών τιμών του εξεταζόμενου δείγματος (Χρήστου, 2007).

Το επόμενο στάδιο της ΠΓΠ είναι η ερμηνεία των αποτελεσμάτων της. Στον Πίνακα Variables Entered/Remove διαπιστώνεται η διαδικασία που ακολουθείται για την εισαγωγή των μεταβλητών στο μοντέλο. Βλέπουμε από τον παρακάτω Πίνακα ότι η πρώτη μεταβλητή και τελικά η μοναδική που συμμετείχε στο μοντέλο είναι η περιφέρεια μηρού (miros). Αξίζει να τονιστεί ότι η διαδικασία κατά την οποία κατασκευάζεται το μοντέλο διακόπτεται αφού εισαχθεί η πρώτη μεταβλητή. Ύστερα από κάθε στάδιο της μεταβλητής που εισάγουμε στο μοντέλο ελέγχεται ως προς τη στατιστική σημαντικότητα της εισαγωγής της μεταβλητής στην εξίσωση της παλινδρόμησης. Επιπλέον σε κάθε στάδια κατασκευής του μοντέλου γίνεται έλεγχος

της νέας μεταβλητής που εισάγεται καθώς και των μεταβλητών που ήδη συμμετέχουν στο μοντέλο. Τα αποτελέσματα από τους ελέγχους συνοψίζονται στον παρακάτω Πίνακα 4.7.

Πίνακας 4.7: Μοντέλο ΠΓΠ

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Περιφέρεια μηρου	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: Δείκτης σωματικού λίπους

Με τη μέθοδο Stepwise που ακολουθείται μας δίνεται η δυνατότητα να εισαχθεί η μεταβλητή και να προσαρμοστεί στο μοντέλο μας. Παρακάτω παρουσιάζεται η σύνοψη του μοντέλου παρουσιάζοντας μερικούς βασικούς δείκτες οι οποίοι δείχνουν την καλή προσαρμογή στο μοντέλο. Ο δείκτης του συντελεστή προσδιορισμού R Square αποτελεί την ένδειξη του ποσοστού της διακύμανσης της εξαρτημένης μεταβλητής που επεξηγεί το μοντέλο δηλαδή δείχνει το ποσοστό στο οποίο ο δείκτης σωματικού λίπους ερμηνεύεται από την ανεξάρτητη μεταβλητή περιφέρεια μηρού. Φαίνεται ότι το 77,1% του δείκτη σωματικού λίπους ερμηνεύεται από την ανεξάρτητη μεταβλητή περιφέρεια μηρού. Αξίζει να αναφερθεί ότι όταν οι τιμές του δείκτη βρίσκονται κοντά στο ένα αντιλαμβάνεται κανείς ότι οι παράγοντες του μοντέλου είναι ικανοποιητικοί για να περιγράψουν τη κίνηση της εξαρτημένης μεταβλητής. Αντίθετα αν οι τιμές του δείκτη βρίσκονται κοντά στο μηδέν βλέπουμε ότι οι ανεξάρτητες μεταβλητές που προτείνονται δεν είναι ικανοποιητικές για να περιγράψει την εξαρτημένη τιμή. Κατά συνέπεια σε αυτή τη περίπτωση το μοντέλο της παλινδρόμησης δεν ενδείκνυται για να προβλεφθούν οι τιμές.

Πίνακας 4.8: Σύνοψη του μοντέλου ΠΓΠ

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,878[a]	,771	,758	2,5102

a. Predictors: (Constant), Περιφέρεια μηρου

b. Dependent Variable: Δείκτης σωματικού λίπους

Στη συνέχεια παρουσιάζεται ο πίνακας της Ανάλυσης Διακύμανσης (ANOBA) ο οποίος απεικονίζει το συνολικό έλεγχο που είναι υπεύθυνος για τη στατιστική σημαντικότητα του μοντέλου παλινδρόμησης. Ο παραπάνω έλεγχος στηρίζεται στη συνάρτηση F ελέγχοντας την υπόθεση ότι οι συντελεστές των ανεξάρτητων μεταβλητών στο μοντέλο είναι ταυτόχρονα μηδέν. Στη περίπτωση που το Sig. είναι μικρότερο του 0.05 τότε η αρχική υπόθεση απορρίπτεται. Από το γεγονός αυτό συμπεραίνουμε ότι το μοντέλο που εξετάζουμε είναι στατιστικά σημαντικό.

Πίνακας 4.9:Ανάλυση Διακύμανσης (ANOBA)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	381,966	1	381,966	60,617	,000[b]
	Residual	113,424	18	6,301		
	Total	495,389	19			

a. Dependent Variable: Δείκτης σωματικού λίπους

b. Predictors: (Constant), Περιφέρεια μηρού

Από το πίνακα της ανάλυσης διακύμανσης εξάγουμε συμπεράσματα σχετικά με τη συνολική διακύμανση του δείγματος που ερμηνεύεται από το μοντέλο παλινδρόμησης. Το συνολικό ποσοστό της διακύμανσης του δείγματος Total= 495,389 αποτελεί το άθροισμα της διακύμανσης της παλινδρόμησης (regression 381,966) και της διακύμανσης του λάθους (Residual = 113,424). Το μεγαλύτερο μέρος της συνολικής διακύμανσης εξηγείται καλύτερα όσο καλύτερο είναι το μοντέλο της παλινδρόμησης. Βλέπουμε ότι ο F έλεγχος είναι 60,617 και το sig=0,000 γεγονός που αποδεικνύει τη στατιστική σημαντικότητα του μοντέλου.

Στη συνέχεια το μοντέλο αναλύεται από τον πίνακα Coefficients. Οι συντελεστές έχουν υπολογιστεί για τα τρία διαδοχικά μοντέλα. Το μοντέλο της γραμμικής παλινδρόμησης γράφεται με την εξής εξίσωση:

Δείκτης σωματικού λίπους = -23,6 + 0,85*Περιφέρεια μηρού (4.9)

Από την παραπάνω εξίσωση και από τον πίνακα των συντελεστών συμπεραίνουμε ότι αν αυξηθεί η ανεξάρτητη μεταβλητή περιφέρεια του μηρού κατά 1 μονάδα, ο δείκτης σωματικού λίπους θα αυξηθεί κατά 85,7%. Αυτή η επίδραση είναι στατιστικά σημαντική όπως διαπιστώνεται και από το sig=0. Αν η περιφέρεια του μηρού έπαιρνε τη τιμή μηδέν τότε με μια αύξηση του σταθερού όρου κατά 1 μονάδα, ο δείκτης του σωματικού λίπους θα σημείωνε μείωση κατά 23,6 μονάδες. Και αυτή η επίδραση είναι στατιστικά σημαντική καθώς το sig του σταθερού όρου είναι 0,001 δηλαδή μικρότερο από 0,05.

Πίνακας 4.10: Συντελεστές παλινδρόμησης

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-23,634	5,657		-4,178	,001			
	Περιφέρεια μηρου	,857	,110	,878	7,786	,000	,878	,878	,878

a. Dependent Variable: Δείκτης σωματικού λίπους

Από το παραπάνω Πίνακα παρατηρούμε τους συντελεστές συσχέτισης της εξίσωσης που εμφανίζονται στη στήλη B. Η στήλη με τη τυπική απόκλιση (Std. Error) περιέχει τις τιμές του τυπικού σφάλματος της εκτίμησης των συντελεστών B. Στη στήλη με το σύμβολο t παρουσιάζονται οι τιμές της στατιστικής συνάρτησης μέσα από την οποία ελέγχεται η σημαντικότητα των συντελεστών της ανάλυσης. Στη στήλη sig., παρατίθεται η τιμή σημαντικότητας με βάση την οποία θα διαπιστωθεί αν θα διατηρηθεί η όχι η μεταβλητή στο μοντέλο.

Πίνακας 4.11: Μοντέλο- Excluded Variables

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Πάχος δέρματος	,219[b]	,733	,474	,175	,147
	Περιφέρεια μπράτσου	,069[b]	,595	,560	,143	,993

a. Dependent Variable: Δείκτης σωματικού λίπους

b. Predictors in the Model: (Constant), Περιφέρεια μηρου

Στον Πίνακα 4.11 βλέπουμε ότι το πάχος του δέρματος που είναι predictor στο μοντέλο 1, δεν είναι στατιστικά σημαντική μεταβλητή καθώς sig=0,474>0,05. Το ίδιο συμβαίνει και με τη μεταβλητή περιφέρεια μπράτσου που λειτουργεί ως predictor, όπου sig=0,560>0,05. Συνεπώς δεν απορρίπτουμε την αρχική μηδενική υπόθεση. Η συσχέτιση και για τις δυο μεταβλητές όπως φαίνεται από τους συντελεστές μερικής συσχέτισης είναι αρνητική. Τέλος φαίνεται ότι δεν υπάρχει στατιστικά σημαντική επίδραση των 2 ανεξάρτητων μεταβλητών (derma, bratso) πάνω στην εξαρτημένη μεταβλητή, γεγονός που επιβεβαιώνεται και από τους συντελεστές beta. Τόσο στη μέθοδο Stepwise όπως και στις μεθόδους Forward και Backward συμπεραίνουμε ότι όλες οι μεταβλητές που συμμετέχουν τελικά στο μοντέλο είναι στατιστικά σημαντικές. Στον Πίνακα 4.12 παρουσιάζονται τα στατιστικά μέτρα των καταλοίπων που είχαμε επιλέξει κατά την αποθήκευση των δεδομένων.

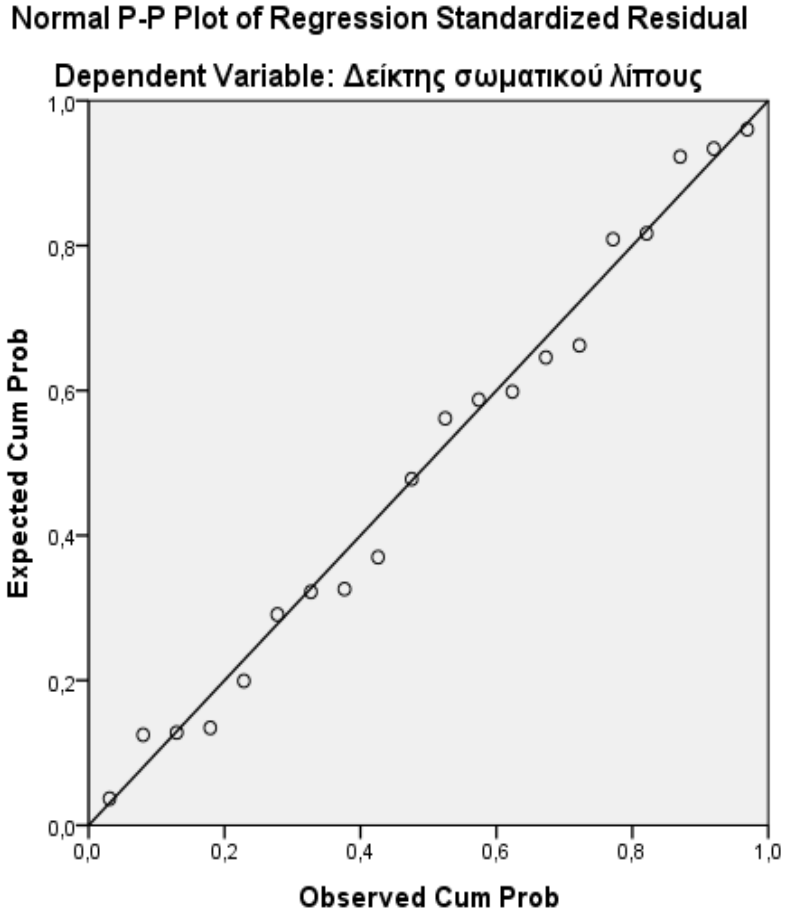
Πίνακας 12: Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	12,512	26,559	20,195	4,4837	20
Std. Predicted Value	-1,714	1,419	,000	1,000	20
Standard Error of Predicted Value	,562	1,135	,771	,195	20
Adjusted Predicted Value	12,412	26,773	20,195	4,4931	20
Residual	-4,4949	4,4084	,0000	2,4433	20
Std. Residual	-1,791	1,756	,000	,973	20
Stud. Residual	-1,879	1,803	,000	1,015	20
Deleted Residual	-4,9498	4,6486	-,0003	2,6601	20
Stud. Deleted Residual	-2,037	1,936	,004	1,054	20
Mahal. Distance	,001	2,936	,950	,975	20
Cook's Distance	,000	,198	,044	,055	20
Centered Leverage Value	,000	,155	,050	,051	20

a. Dependent Variable: Δείκτης σωματικού λίπους

Το τελευταίο στάδιο μετά τον υπολογισμό του μοντέλου της γραμμικής παλινδρόμησης είναι ο έλεγχος της τελευταίας υπόθεσης της κανονικότητας των καταλοίπων. Επιβεβαιώνονται οι τυχαίες κινήσεις που λαμβάνουν χώρα μέσα από το διάγραμμα σημείων της εξαρτημένης μεταβλητής με τις τιμές των λαθών.

Διάγραμμα 4.1: Διάγραμμα σημείων του δείκτη σωματικού λίπους-των καταλοίπων



Κεφάλαιο 5: Έλεγχος ανεξαρτησίας ποιοτικών μεταβλητών

5.1 Στατιστικοί έλεγχοι υποθέσεων

Οι στατιστικοί έλεγχοι των ερευνητικών υποθέσεων πραγματοποιούνται με διάφορους στατιστικούς ελέγχους, οι οποίοι εξετάζουν αν υπάρχει ή όχι στατιστικά σημαντική διαφορά μεταξύ ενός, δύο ή περισσοτέρων δειγμάτων αν τα δείγματα αυτά αντιπροσωπεύουν διαφορετικούς πληθυσμούς (Ζαχαροπούλου, 2010).

Ο έλεγχος της ανεξαρτησίας δυο ποιοτικών μεταβλητών επιτυγχάνεται με τον στατιστικό έλεγχο X^2 τον οποίο αναλύουμε παρακάτω. Με τον όρο ποιοτικές μεταβλητές εννοούμε τις μεταβλητές που δεν μετρώνται ποσοτικά. Δείχνουν ότι οι διάφοροι παράγοντες μεταβάλλονται κατά είδος. Παραδείγματα ποιοτικών μεταβλητών αποτελούν το φύλο", το "χρώμα των ματιών", το "βάρος", το "ύψος", η "κοινωνική κατάσταση" κ.α. Εκείνες που παρέχουν τη δυνατότητα διάταξης ονομάζονται διατάξιμες (π.χ. η διαγωγή ενός μαθητή) (Παπαϊωάννου & Λουκάς, 2002). Σύμφωνα με τη βιβλιογραφία υπάρχει ένα πλήθος στατιστικών μέτρων τα οποία είναι διαθέσιμα ανάλογα με τη φύση των μεταβλητών και χρησιμοποιούνται για να καθοριστεί η ένταση της σχέσης μεταξύ των δύο ποιοτικών μεταβλητών[6].

Οι στατιστικοί έλεγχοι υποθέσεων όπως ήδη έχουμε αναφέρει χρησιμοποιούνται για τον έλεγχο της κανονικότητας των μεταβλητών που χρησιμοποιούμε. Ο στατιστικός έλεγχος υποθέσεων (hypothesis testing) αποτελεί μια συμπερασματική διαδικασία/μέθοδος που προσφέρει η Στατιστική Συμπερασματολογία και εφαρμόζεται σε στοχαστικά προβλήματα απόφασης μεταξύ δύο εναλλακτικών υποθέσεων. Οι υποθέσεις που κάνουμε είναι η μηδενική και η εναλλακτική υπόθεση. Η μηδενική υπόθεση (H_0 null hypothesis) είναι η εξής: η υπό έλεγχο κατανομή, δε διαφέρει από την κανονική κατανομή έναντι της εναλλακτικής υπόθεσης (H_1 alternative hypothesis) η οποία είναι η εξής: η υπό έλεγχο κατανομή διαφέρει από την κανονική κατανομή (Montgomery, et al., 2012).

Όταν μελετάμε τις ποιοτικές μεταβλητές η μηδενική υπόθεση που χρησιμοποιούμε είναι η εξής: οι παρατηρηθείσες συχνότητες είναι ίσες με τις αναμενόμενες συχνότητες (δεν υπάρχει σχέση ανάμεσα στις δύο εξεταζόμενες μεταβλητές). Η εναλλακτική υπόθεση είναι: οι παρατηρηθείσες συχνότητες και οι αναμενόμενες συχνότητες διαφέρουν (υπάρχει σχέση) (Ιωαννίδης, 2005).

Βασική προϋπόθεση προκειμένου να εφαρμοστούν σωστά οι στατιστικοί έλεγχοι υποθέσεων είναι να ερμηνευτούν σωστά τα αποτελέσματα και να κατανοηθεί απόλυτα η λογική και το νόημά τους. Για να γίνουμε πιο κατανοητοί αξίζει να σημειωθεί ότι κατά το στατιστικό έλεγχο υποθέσεων ο αναλυτής θέτει σα μηδενική υπόθεση (H_0) εκείνη για την οποία έχει αμφιβολίες δηλαδή αυτή που αμφισβητείται περισσότερο και εξετάζει την περίπτωση που το τυχαίο δείγμα που έχει επιλέξει από έναν τυχαίο πληθυσμό μας κατευθύνει προς την απόρριψη της αρχική υπόθεσης έναντι της εναλλακτικής υπόθεσης (H_1). Δηλαδή, η H_0 απορρίπτεται ή δεν απορρίπτεται με βάση το τι παρατηρείται στο τυχαίο δείγμα που πήραμε από τον πληθυσμό (Champkin, 2013).

[6] βλέπε σχετικά Παπαϊωάννου και Λουκάς, 2002, σελ. 289-292, Παπαϊωάννου και Φερεντίνος, 2000, σελ. 270-276

Πιο συγκεκριμένα, υποθέτοντας ότι η H_0 είναι αληθής, αν αυτό που παρατηρείται στο δείγμα είναι ακραίο, δηλαδή, αν έχει πολύ μικρή πιθανότητα να συμβεί, τότε απορρίπτουμε την H_0. Σε αντίθετη περίπτωση, δηλαδή, αν αυτό που παρατηρείται στο δείγμα δεν είναι ακραίο-σπάνιο (όταν είναι αληθής η H_0) τότε το δείγμα που πήραμε δε μας δίνει αρκετές ενδείξεις για την απόρριψη της H_0 και «αποτυγχάνουμε να την απορρίψουμε» (Κικιλίας, et al., 2001).

Βέβαια, ακολουθώντας αυτή τη στρατηγική λαμβάνουμε υπόψη ότι υπάρχει πιθανότητα να συμβούν ακόμα και τα «ακραία» έστω και με πολύ μικρή πιθανότητα. Στην περίπτωση που απορρίπτουμε λανθασμένα την H_0 τότε έχουμε κάνει σφάλμα το οποίο ονομάζεται σφάλμα τύπου I (type I error). Εφόσον, υπό την H_0, υπάρχει πιθανότητα έστω πολύ μικρή (π.χ. 0.0001), το ακραίο να συμβεί τότε απορρίπτουμε λανθασμένα την H_0 με πιθανότητα 0.0001. Ανάλογα, είναι δυνατόν, λανθασμένα να μην απορρίψουμε την H_0. Δηλαδή, να αποτύχουμε να απορρίψουμε την H_0, ενώ είναι αληθής η H_1. Αυτό το σφάλμα ονομάζεται σφάλμα τύπου II (type II error). Το «ρίσκο», επομένως, είναι διπλό, με πιθανότητα λανθασμένης απόρριψης της H_0 δηλαδή P(σφάλμα τύπου I) = P(απόρριψη της H_0 | αληθής η H_0) και λανθασμένης μη απόρριψης της H_0 δηλαδή P (σφάλμα τύπου II) = P(μη απόρριψη της H_0 | αληθής η H_1) (Κιντής, 2010).

Κατά κύριο λόγο η H_0 δείχνει ότι μια κατάσταση δεν μεταβάλλεται δηλαδή ότι η ανεξάρτητη μεταβλητή δεν ασκεί καμία επίδραση στην εξαρτημένη μεταβλητή για τον πληθυσμό που εξετάζουμε. Επιπλέον για να καθορίσουμε την υπόθεση H_0 υπάρχει και άλλος ένας τρόπος σύμφωνα με τον οποίο μπορούμε να ορίσουμε ως μηδενική υπόθεση την υπόθεση την οποία αν ορίσουμε λάθος θα έχουμε περισσότερο κίνδυνο. Δηλαδή θέτουμε ως μηδενική υπόθεση την υπόθεση που χρειάζεται να προστατευθεί περισσότερο από το σφάλμα τύπου I (Κιντής, 2010).

Ενώ η εναλλακτική υπόθεση H_1 υποδηλώνει ότι στο δείγμα που εξετάζουμε υπάρχει μεταβολή που σημαίνει ότι η ανεξάρτητη μεταβλητή ασκεί επίδραση στην εξαρτημένη για τον εξεταζόμενο πληθυσμό. Οι δυο έλεγχοι που εντοπίζονται είναι ο ένας μονόπλευρος και ο άλλος δεξιόπλευρος όπως φαίνονται παρακάτω:

$$H_0: \mu \leq \mu_0 \quad H_0: \mu \geq \mu_0$$

$$H_1: \mu > \mu_0 \quad H_1: \mu < \mu_0$$

Ονομάζονται μονόπλευροι έλεγχοι (δεξιόπλευρος και αριστερόπλευρος αντίστοιχα) ενώ ο έλεγχος,

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

ονομάζεται αμφίπλευρος. Αξίζει να σημειωθεί επίσης ότι τα δύο σύνολα τιμών της παραμέτρου που ελέγχουμε από τις δυο υποθέσεις, πρέπει να είναι ξένα μεταξύ τους ή το ένα να αρνείται το άλλο (Χάλκος, 2011). Στη συνέχεια ορίζουμε ένα μέγιστο ανεκτό μέγεθος σφάλματος τύπου I το οποίο συμβολίζεται με α και ονομάζεται επίπεδο στατιστικής σημαντικότητας (level of significance) του ελέγχου (γιατί από αυτό προκύπτει η τιμή που ορίζει αν αυτό που παρατηρείται στο δείγμα είναι σημαντικό-σημαντική απόδειξη για να υποστηρίξει την απόρριψη της H_0).

Συνήθως το επίπεδο σημαντικότητας α ορίζεται ίσο με 0.01 ή 0.05. Αν η τιμή P-value είναι μικρότερη από 0.05 τότε απορρίπτουμε τη μηδενική υπόθεση, αν είναι

μεγαλύτερη από 0.05 δεν απορρίπτουμε τη μηδενική υπόθεση. Αν υπολογίσουμε τη τιμή P-value στην ανάλυση μας συμπεραίνουμε την πιθανότητα εμφάνισης του δείγματος που εξετάζουμε υπό την υπόθεση ότι η μηδενική είναι αληθής. Συνεπώς όσο μικρότερη είναι η τιμή P-value τόσο πιο ισχυρές ενδείξεις έχουμε εναντίον της μηδενικής υπόθεσης δηλαδή είναι πολύ σημαντική η τιμή της στατιστικής συνάρτησης που δίνει το εξεταζόμενο δείγμα. Με άλλα λόγια η τιμή P-value δείχνει τη πιθανότητα να έχουμε κάνει λάθος. Οι τιμές P-value ελέγχουν την υπόθεση ότι κάθε συντελεστής είναι διαφορετικός του μηδενός. Για να συμβεί αυτό η τιμή P-value πρέπει να είναι μικρότερη του 0,05. Στη περίπτωση που απορρίπτουμε τη μηδενική υπόθεση H_0, τότε χαρακτηρίζουμε το δείγμα που εξετάζουμε ως στατιστικά σημαντικό (statistically significant), γεγονός που σημαίνει ότι έχει σημαντική διαφορά από την αναμενόμενη τιμή της μηδενικής υπόθεσης (Χρήστου, 2007).

Είναι αξιοσημείωτο ότι αν θέσουμε πιο μικρό επίπεδο σημαντικότητας τότε είναι φανερό ότι επιδιώκουμε να έχουμε ένα στατιστικά σημαντικό δείγμα και να απορρίψουμε την μηδενική υπόθεση ώστε να επιτύχουμε το σκοπό μας. Συνεπώς υπάρχει πιθανότητα σε κάποιο επίπεδο στατιστικής σημαντικότητας α, για παράδειγμα για α=0.05 απορρίπτουμε τη μηδενική υπόθεση H_0 ενώ για α=0.01 να μην απορρίπτουμε την μηδενική υπόθεση γιατί επιδιώκουμε το δείγμα μας να είναι στατιστικά σημαντικό. Το επίπεδο σημαντικότητας στο οποίο απορρίπτεται η μηδενική υπόθεση είναι συνήθως μικρό και όσο πιο μικρό είναι τόσο πιο ισχυρές αποδείξεις δημιουργούνται έναντι της μηδενικής υπόθεσης. Κατά συνέπεια το αποτέλεσμα ελέγχου είναι στατιστικά σημαντικό για το δείγμα μας στην περίπτωση που έχουμε μικρό επίπεδο σημαντικότητας στο οποίο απορρίπτουμε τη μηδενική υπόθεση. Τέλος γίνεται φανερό ότι στη περίπτωση που η μηδενική υπόθεση H_0 απορριφθεί σε κάποιο επίπεδο σημαντικότητας α τότε απορρίπτεται και σε μεγαλύτερο επίπεδο σημαντικότητας (Κιντής, 2010). Όλα τα παραπάνω παρουσιάζονται με παραδείγματα σε επόμενες ενότητες.

5.2 Έλεγχος Χ-τετράγωνο

Ο έλεγχος Χ-τετράγωνο χρησιμοποιείται προκειμένου να διαπιστωθεί η συσχέτιση μεταξύ δύο κατηγορικών ή διατεταγμένων μεταβλητών. Ο σκοπός του σε μια έρευνα είναι να παρέχει την απαραίτητη πληροφόρηση στον ερευνητή για την ένταση της συσχέτισης μεταξύ των μεταβλητών. Αυτό που δεν έχει τη δυνατότητα να παρέχει είναι μια ένδειξη για την κατεύθυνση της συσχέτισης (Γναρδέλλης, 2003). Στην περίπτωση της ύπαρξης πινάκων διπλής κατεύθυνσης με μεταβλητές που έχουν δύο κατηγορίες η κάθε μία (πίνακας 2x2) γίνεται χρήση του Fisher's exact test, το οποίο ακολουθεί τη λογική του ελέγχου x^2 εφαρμοζόμενο αποκλειστικά σε πίνακες 2x2 (Σιώμκος & Βασιλακοπούλου, 2005).

Για να βρούμε την πιθανή σχέση μεταξύ δύο ποιοτικών μεταβλητών δημιουργούμε το πίνακα συνάφειας όπως θα δούμε παρακάτω, ο οποίος περιέχει γραμμές που αντιστοιχούν στις κατηγορίες της μιας μεταβλητής που εξετάζουμε και συμβολίζεται με r. Οι στήλες του πίνακα τις συμβολίζουμε με το c και μας δείχνουν τις κατηγορίες της άλλης ποιοτικής μεταβλητής στο δείγμα μας. Έτσι έχουμε το πολλαπλασιασμό r x c κελιά. Κάθε κελί αποτελεί έναν συνδυασμό αυτών των δυο μεταβλητών. Σε αυτά τα κελιά παρουσιάζονται οι συχνότητες εμφάνισης αυτών των μεταβλητών. Προκειμένου να ελέγξουμε αν υπάρχει η όχι ανεξαρτησία ανάμεσα σε δυο ποιοτικές μεταβλητές χρησιμοποιούμε τον έλεγχο $\mathbf{X^2}$ το οποίο αποτελεί ένα στατιστικό τεστ και υπολογίζεται από την παρακάτω σχέση:

$$X^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij} - E_{ij})^2}{E_{ij}} \quad \textbf{(5.1)}$$

Όπου

- O_{ij} , η παρατηρούμενη συχνότητα του (i, j) κελιού δηλαδή i , j οι δυο ποιοτικές μεταβλητές
- E_{ij} , η συχνότητα που αναμένουμε να πάρουμε σε αυτό το κελί στην περίπτωση που οι μεταβλητές που εξετάζουμε είναι ανεξάρτητες στατιστικά.

Η E_{ij} υπολογίζεται από την εξής σχέση:

$$E_{ij} = \frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}} = \frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}}{n} \quad \textbf{(5.2)}$$

όπου n το μέγεθος του δείγματος. Αντιλαμβάνεται κανείς ότι όταν παρατηρείται μεγάλη απόκλιση των αναμενόμενων τιμών από τις παρατηρούμενες υπάρχει μια πιθανότητα να έχουμε μια σχέση εξάρτησης. Η υπόθεση που έχουμε κάνει σχετικά με την ανεξαρτησία των μεταβλητών απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας α (συνήθως ορίζεται α=0.05), όταν:

$$X^2 \geq X^2_{(r-1)(c-1),a} \quad \textbf{(5.3)}$$

Ή όταν p-value <α. Στην περίπτωση που η μηδενική υπόθεση που έχουμε κάνει δηλαδή η υπόθεση ανεξαρτησίας απορρίπτεται τότε προχωράμε στα επόμενα βήματα.

Αξίζει να σημειωθεί στο σημείο αυτό πριν προχωρήσουμε στην ανάλυση των επόμενων βημάτων ότι το παραπάνω τεστ εφαρμόζεται υπό κάποιες προϋποθέσεις. Η πρώτη υπόθεση είναι το μέγεθος του δείγματος n, να είναι το τετραπλάσιο σε σχέση με το πλήθος των κελιών και η δεύτερη υπόθεση είναι οι αναμενόμενες συχνότητες να μην είναι μικρότερες της μονάδας καθώς και το 25% εξ' αυτών να μην είναι μικρότερες από το 5. Στη περίπτωση που δεν πληρούνται αυτές οι δυο προϋποθέσεις τότε στην περίπτωση που έχουμε 2X2 κελιά χρησιμοποιούμε το ακριβές στατιστικό του Fisher. Σε κάθε άλλη περίπτωση είναι απαραίτητη η συγχώνευση των γειτονικών κελιών με τρόπο ώστε να εξαλείφεται το παραπάνω πρόβλημα αλλά ταυτόχρονα να υπάρχει φυσική ερμηνεία των νέων κατηγοριών-κελιών. Η συγχώνευση των κελιών επιτυγχάνεται με επανακωδικοποίηση (recode) μίας εκ των δύο ποιοτικών μεταβλητών. Επιπλέον στη περίπτωση 2X2 πινάκων χρησιμοποιείται η διόρθωση συνεχείας του Yates (Continuity Correction) αντί του κλασικού X-τετράγωνο τεστ (Σιώμκος & Βασιλακοπούλου, 2005).

Το τρίτο βήμα περιλαμβάνει τη διερεύνηση της έντασης και της φύσης της σχέσης μεταξύ των δύο εξεταζόμενων μεταβλητών. Για να επιτευχθεί αυτό υπάρχει μια πληθώρα στατιστικών μέτρων που μπορούμε να χρησιμοποιήσουμε. Κάποια από αυτά τα στατιστικά μέτρα είναι τα εξής:

α) Ο συντελεστής συνάφειας ή σύμπτωσης (contigency coefficient),

$$C = \sqrt{\frac{X^2}{(X^2+n)}} \quad \textbf{(5.4)}$$

Ο παραπάνω συντελεστής όταν παίρνεις τιμές κοντά στο 0 πρόκειται για ανεξάρτητες μεταβλητές ενώ η μέγιστη τιμή που μπορεί να πάρει είναι μικρότερη της μονάδας 1. Όμως αυτό εξαρτάται από τον αριθμό των κατηγοριών των δύο ποιοτικών μεταβλητών.

β) ο συντελεστές που ήδη έχουμε εξετάσει, Phi (συντελεστής του Pearson),

$$\boldsymbol{\Phi} = \sqrt{\frac{X^2}{n}} \textbf{ (5.5)}$$

η μέγιστη τιμή που παίρνει εξαρτάται από το μέγεθος του πίνακα. Η μηδενική τιμή που λαμβάνει αποδεικνύει την ανεξαρτησία των μεταβλητών.

γ) ο συντελεστής V του Cramer,

$$\boldsymbol{V} = \sqrt{\frac{X^2}{nmin(r-1,c-1)}} \textbf{ (5.6)}$$

Ο συγκεκριμένος συντελεστής ταυτίζεται στην περίπτωση των 2Χ2 πινάκων με το συντελεστή Phi ενώ οι τιμές που παίρνει είναι από 0 που δηλώνει ανεξαρτησία έως 1 που δηλώνει απόλυτη συνάφεια.

δ) ο συντελεστής Lambda ανομάζεται και *Goodman-Kruskal lambda* και οι συντελεστές αβεβαιότητας (uncertainty coefficient) οι οποίοι ονομάζονται και Theil's U.

Στην περίπτωση που θέλουμε να ελέγξουμε τη σχέση μεταξύ διατάξιμων (Ordinal) ποιοτικών μεταβλητών έχουμε τη δυνατότητα να χρησιμοποιήσουμε στατιστικά μέτρα (δηλαδή συντελεστές) που μπορούν να προσδιορίσουν τη φύση της συνάφειας (θετικής ή αρνητικής). Τα εν λόγω μέτρα παίρνουν τιμές στο διάστημα [-1,1]. Όταν λαμβάνουν τη τιμή -1 σημαίνει ότι έχουμε τέλεια αρνητική συνάφεια ενώ όταν λαμβάνουν τη τιμή 0 οδηγούμαστε σε μη ύπαρξη συνάφειας, η τιμή 1 υποδηλώνει τέλεια θετική συνάφεια. Ανάμεσα σε πολλά στατιστικά μέτρα μπορούμε να χρησιμοποιήσουμε και άλλους συντελεστές όπως είναι ο Gamma, Kendall’s tau-b ο οποίος ενδείκνυται για συμμετρικούς πίνακες, Kendall’s tau-c ο οποίος ενδείκνυται για μη συμμετρικούς πίνακες και ο Somers’ d ο οποίος ενδείκνυται για τη περίπτωση που η μια από τις δυο μεταβλητές είναι εξαρτημένη και η άλλη ανεξάρτητη (Μπατσίδης, 2014).

Επιπρόσθετα όταν έχουμε μια ποιοτική μεταβλητή που είναι ονοματική και η άλλη διαστηματική χρησιμοποιούμε το συντελεστή Eta ο οποίος παίρνει τιμές στο διάστημα [0,1]. Όταν λαμβάνει τη τιμή 0 υποδηλώνει τη μη ύπαρξη σχέσης, ενώ όταν λαμβάνει τη τιμή 1 υποδεικνύει την ύπαρξη υψηλού βαθμού σχέσης (Μπατσίδης, 2014).

Τέλος, ο συντελεστής Kappa του Kohen μπορεί να χρησιμοποιηθεί στους πίνακες συνάφειας οι οποίες περιλαμβάνουν τις ίδιες κατηγορίες τόσο στις στήλες όσο και στις γραμμές. Οι τιμές που παίρνει βρίσκονται στο διάστημα [-1,1]. Η τιμή 1 (-1 αντίστοιχα) δηλώνει τη πλήρη συμφωνία (πλήρη διαφωνία αντίστοιχα) μεταξύ των μεταβλητών, ενώ η τιμή 0 υποδεικνύει ότι η συμφωνία είναι τυχαία (Μπατσίδης, 2014).

5.3 Εξέταση της σχέσης μεταξύ ποιοτικών μεταβλητών με χρήση του SPSS

Για να εξετάσουμε τη σχέση μεταξύ ποιοτικών μεταβλητή με τη χρήση του στατιστικού πακέτου SPSS επιλέγουμε ένα τυχαίο δείγμα 35 παιδιών προσχολικής ηλικίας (Ζωγράφος & Γναρδέλλης, 2003). Οι ποιοτικές μεταβλητές του δείγματος είναι οι εξής: το φύλο, η διαγωγή, η οικονομική κατάσταση της οικογένειας. Τα δεδομένα του δείγματος που αφορούν τις συγκεκριμένες μεταβλητές παρουσιάζονται στην Εικόνα 5.1. Συγκεκριμένα στη στήλη που περιλαμβάνει το Φύλο έχουμε για Α= Αγόρι και Θ= Κορίτσι, στη στήλη με τη λέξη Διαγωγή έχουμε Α= Κοσμιωτάτη και Β= Κοσμία και στη στήλη που δείχνει την Οικονομική Κατάσταση έχουμε Α=0-450, Β= 450-600, Γ=600-900 και Δ= 900 ευρώ και άνω (Μπατσίδης, 2014).

Εικόνα 5.1: Ποιοτικές μεταβλητές δείγματος

Φύλο	Διαγωγή	Οικον. Κατάσταση
Α	Β	Β
Θ	Α	Γ
Θ	Α	Γ
Θ	Α	Γ
Α	Α	Α
Α	Α	Β
Θ	Β	Β
Α	Α	Δ
Α	Α	Γ
Α	Α	Β
Α	Α	Β
Θ	Α	Α
Θ	Α	Γ
Α	Α	Δ
Θ	Α	Β
Θ	Β	Β
Α	Α	Α
Α	Α	Γ
Α	Α	Γ
Α	Α	Α
Α	Α	Β
Θ	Α	Δ
Θ	Α	Δ
Α	Α	Β
Θ	Α	Β
Θ	Α	Α
Α	Β	Α
Α	Α	Γ
Α	Α	Γ
Θ	Α	Α
Θ	Α	Δ
Α	Β	Β
Θ	Α	Α
Α	Α	Γ
Θ	Α	Δ

Για να εισάγουμε τα δεδομένα των ποιοτικών μεταβλητών στο SPSS θα πρέπει αρχικά να τα διαχωρίσουμε ανάλογα με τη μορφή τους είτε σε αριθμητικά δεδομένα (γεγονός που είναι επικρατέστερο) είτε σε χαρακτήρες. Αφού διαχωρίσουμε τα δεδομένα μας προχωράμε στην εισαγωγή των δεδομένων στο πρόγραμμα. Όταν η μορφή των δεδομένων μας είναι αριθμητική απαραίτητη προϋπόθεση είναι η προεργασία των μεταβλητών αυτών. Αυτή η προεργασία περιέχει την αντιστοίχιση των αριθμητικών τιμών σε κάθε πιθανή κατηγορία κάθε ποιοτικής μεταβλητής που

περιλαμβάνεται στο δείγμα μας. Με την κωδικοποίηση καταφέρνουμε να αντιστοιχίζουμε κάθε ποιοτική μεταβλητή σε έναν κωδικό. Κατόπιν καταχωρούμε τους κωδικούς αυτούς σε κάθε κελί. Η διαδικασία που ακολουθείται για την εισαγωγή των ποιοτικών μεταβλητών είναι η ίδια διαδικασία που ακολουθείται κατά την καταχώρηση των ποσοτικών δεδομένων. Για να κωδικοποιήσουμε λοιπόν τις μεταβλητές μας θέτουμε τη μεταβλητή Φύλο ότι θα παίρνει τη τιμή 1 στην περίπτωση που αναφερόμαστε σε Α=Αγόρι και την τιμή 2 όταν πρόκειται για Θ=Κορίτσι. Ακολουθώντας την ίδια λογική θέτουμε τη τιμή 1 στη περίπτωση που έχουμε διαγωγή Α και τη τιμή 2 όταν έχουμε Διαγωγή Β, πρόκειται δηλαδή για μια δυαδική μεταβλητή όπως είναι και η μεταβλητή Φύλο. Όσον αφορά την μεταβλητή Οικονομική Κατάσταση εισάγουμε τις εξής τιμές: 1=Α, 2=Β, 3=Γ και 4=Δ. Τα δεδομένα που προκύπτουν στο S.P.S.S. (τμήμα αυτών) παρουσιάζονται στην Εικόνα 5.2 (Μπατσίδης, 2014).

Εικόνα 5.2: Κωδικοποίηση ποιοτικών μεταβλητών

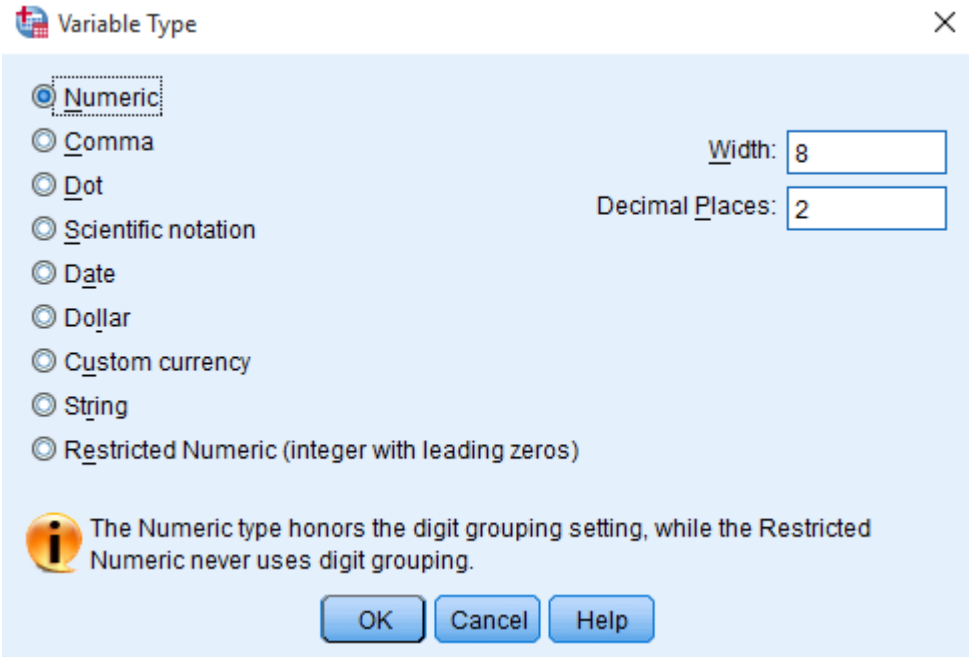
	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006
1	1,00	2,00	2,00	111,00	95,00	22,00
2	2,00	1,00	3,00	90,00	98,00	25,00
3	2,00	1,00	3,00	90,00	92,00	18,00
4	2,00	1,00	3,00	90,00	104,00	19,00
5	1,00	1,00	1,00	104,00	85,00	21,00
6	1,00	1,00	2,00	72,00	96,00	20,00
7	2,00	2,00	2,00	105,00	89,00	21,00
8	1,00	1,00	4,00	93,00	103,00	22,00
9	1,00	1,00	3,00	99,00	110,00	18,00
10	1,00	1,00	2,00	93,00	85,00	27,00
11	1,00	1,00	2,00	84,00	94,00	30,00
12	2,00	1,00	1,00	95,00	98,00	21,00
13	2,00	1,00	3,00	93,00	96,00	24,00
14	1,00	1,00	4,00	78,00	99,00	26,00
15	2,00	1,00	2,00	106,00	83,00	19,00
16	2,00	2,00	2,00	100,00	87,00	27,00
17	1,00	1,00	1,00	81,00	85,00	25,00
18	1,00	1,00	3,00	77,00	97,00	24,00
19	1,00	1,00	3,00	67,00	96,00	23,00
20	1,00	1,00	1,00	100,00	107,00	28,00
21	1,00	1,00	2,00	104,00	102,00	29,00
22	2,00	1,00	4,00	111,00	105,00	19,00
23	2,00	1,00	4,00	122,00	95,00	30,00
24	1,00	1,00	2,00	99,00	82,00	28,00
25	2,00	1,00	2,00	106,00	94,00	31,00
26	2,00	1,00	1,00	126,00	90,00	19,00
27	1,00	2,00	1,00	90,00	90,00	23,00
28	1,00	1,00	3,00	110,00	96,00	32,00
29	1,00	1,00	3,00	117,00	87,00	27,00
30	2,00	1,00	1,00	119,00	97,00	24,00
31	2,00	1,00	4,00	105,00	90,00	22,00
32	1,00	2,00	2,00	100,00	107,00	18,00

Όπως παρατηρούμε από την Εικόνα 5.2 στο S.P.S.S. οι μεταβλητές ονομάζονται VAR00001, VAR00002. Συνεπώς καλούμαστε να μετονομάσουμε τις μεταβλητές στο δείγμα μας. Ακολουθώντας κάποιους βασικούς κανόνες μετονομάζουμε τις μεταβλητές μας με το όνομα που επιθυμούμε και σχετίζεται με τη μεταβλητή που εξετάζουμε. Στο παράθυρο που εμφανίζεται στο δεύτερο φύλλο εργασίας του SPSS και ονομάζεται Variable View και συγκεκριμένα στο πλαίσιο Name εισάγουμε τα ονόματα των μεταβλητών μας τα οποία έχουμε κωδικοποιήσει. Εδώ επιλέγουμε να ονομάσουμε τις μεταβλητές Φύλο, Διάγνωση, Οικονομική κατάσταση ως εξής: Sex, Diagogi, Status αντίστοιχα. Σύμφωνα με τον Μπατσίδη (2014): «στο πεδίο Label δηλώνεται η πλήρης περιγραφή του ονόματος της μεταβλητής που βοηθά στην καλύτερη παρουσίαση των αποτελεσμάτων μας. Με αυτό τον τρόπο δηλώνουμε την ονομασία που θα εμφανίζεται στους πίνακες των

αποτελεσμάτων των αναλύσεων της έρευνας μας» π.χ. Φύλο, Διαγωγή, Οικον. Κατάσταση (Ζωγράφος & Γναρδέλλης, 2003; Μπατσίδης, 2014).

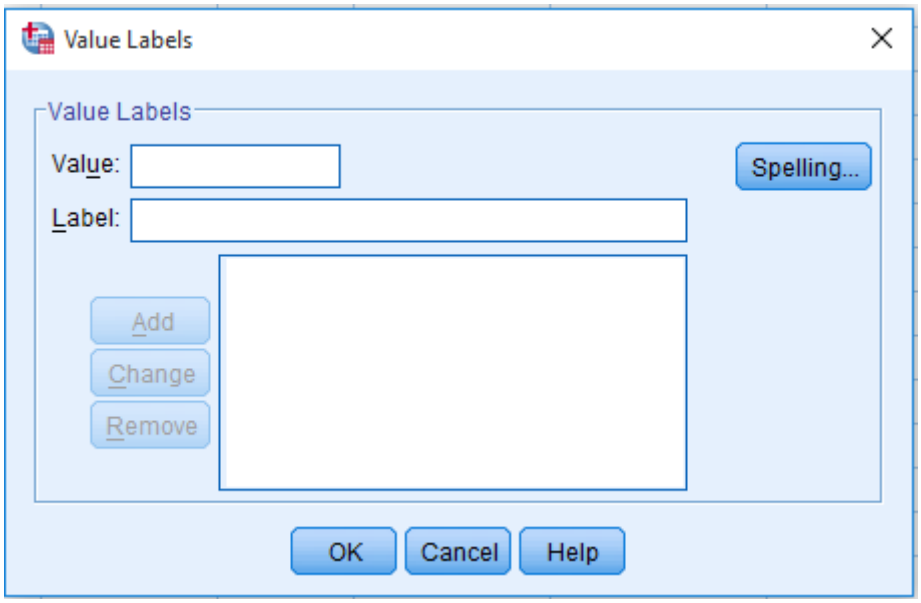
Ένας άλλος τρόπος να καθορίσουμε τον τύπο της μεταβλητής είναι μέσα από το πλαίσιο Variable Type. Ο Μπατσίδης (2014) επισημαίνει ότι: «το πλαίσιο Variable Type βασίζεται στις τιμές που πληκτρολογούμε και καθορίζει αυτόματα τον τύπο της μεταβλητής, έχοντας ως προεπιλογή να τις εμφανίζει αριθμητικές (numeric) με 2 δεκαδικά ψηφία (Decimals Places) και συνολικό μήκος (δηλώνεται στο πλαίσιο Width) 8 θέσεων. Για τον υπολογισμό του μήκους μίας μεταβλητής λαμβάνονται υπόψη το πρόσημο, το ακέραιο μέρος, η δεκαδική τελεία καθώς και το δεκαδικό μέρος της». Αν τα δεδομένα μας είναι τέτοια ώστε να παραβιάζονται αυτές οι προεπιλογές πρέπει να τις τροποποιήσουμε κατάλληλα (Μπατσίδης, 2014).

Εικόνα 5.3: Κωδικοποίηση τύπου μεταβλητής



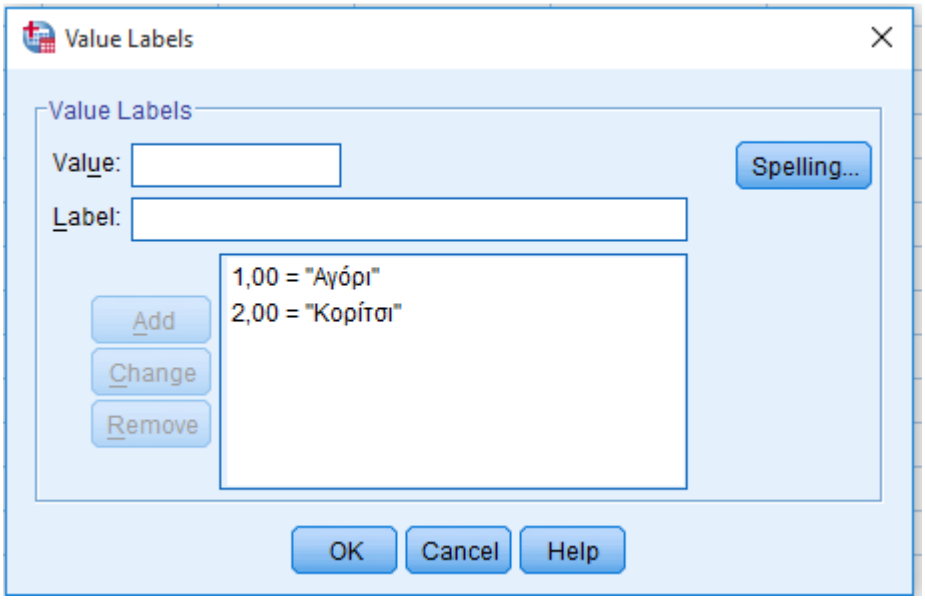
Επόμενο βήμα είναι να θέσουμε τις ετικέτες των τιμών των μεταβλητών που χρησιμοποιούμε. Είναι εύλογο και χρήσιμο για τον ίδιο τον ερευνητή η καταγραφή όλων των κωδικών των μεταβλητών που μελετάμε στο πλαίσιο Values του παραθύρου Variable View. Συνεπώς σε αυτό το πλαίσιο εισάγουμε κατά βάση τις συμβάσεις που πραγματοποιήσαμε κατά τη διαδικασία καταχώρησης των δεδομένων μας. Το γεγονός αυτό γίνεται με το πάτημα του κάτω δεξιού άκρου του κελιού το οποίο σχηματίζεται από την μεταβλητή αλλά και τη στήλη Values. Κατόπιν εμφανίζεται το παράθυρο της Εικόνας 5.4.

Εικόνα 5.4: Ετικέτες τιμών των μεταβλητών



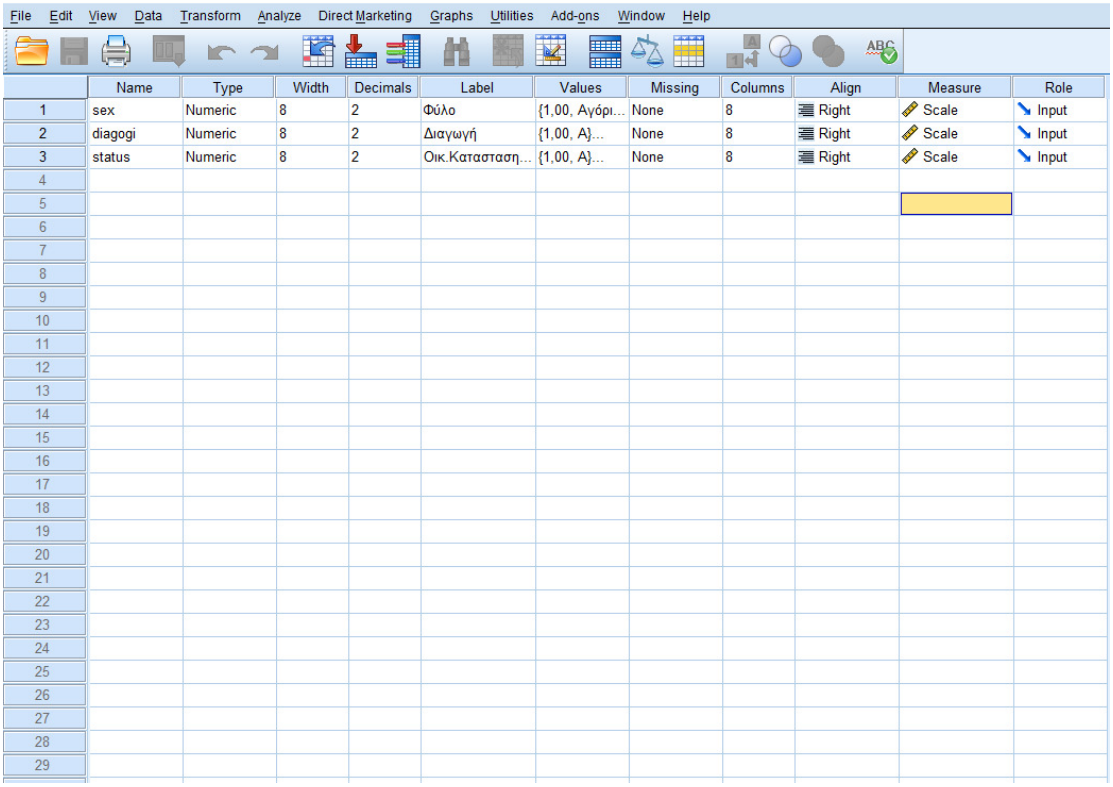
Στο κουτί Value εισάγουμε τη τιμή 1, 2 κλπ και το κουτί Label εισάγουμε τη μεταβλητή ενδιαφέροντος. Για παράδειγμα Value 1, Label Αγόρι, Value 2, Label Κορίτσι. Με την επιλογή του πλήκτρου Add επαναλαμβάνουμε την παραπάνω διαδικασία μέχρι να προστεθεί κάθε μια δυνατή τιμή και ονομασία της κατηγορικής μεταβλητής. Από τη στιγμή που θα ολοκληρωθεί η παραπάνω διαδικασία πατάμε το πλήκτρο OK.

Εικόνα 5.5: Εισαγωγή τιμών



Ακολουθούμε την παραπάνω διαδικασία για όλες τις ποιοτικές μεταβλητές του δείγματός μας. Όταν ολοκληρωθεί η διαδικασία έχουμε το παρακάτω αποτέλεσμα όπως φαίνεται στην Εικόνα 5.6 όπου περιλαμβάνονται μόνο οι ποιοτικές μεταβλητές του δείγματος και έχουμε αφαιρέσει τις ποσοτικές μεταβλητές.

Εικόνα 5.6: Καταχώρηση ποιοτικών μεταβλητών του δείγματος



File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	sex	Numeric	8	2	Φύλο	{1,00, Αγόρι...	None	8	Right	Scale	Input
2	diagogi	Numeric	8	2	Διαγωγή	{1,00, A}...	None	8	Right	Scale	Input
3	status	Numeric	8	2	Οικ.Καταστασ...	{1,00, A}...	None	8	Right	Scale	Input
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											

Το επόμενο σημαντικό βήμα που ακολουθούμε είναι να ελέγξουμε ως προς τις τιμές που λείπουν από τα δεδομένα μας. Αυτή η δυνατότητα μας δίνεται από το παράθυρο Variable View όπου εμφανίζεται το πλαίσιο Missing Values. Σε αυτό το πλαίσιο μπορούμε να καθορίσουμε τις τιμές των ελλιπών τιμών μίας μεταβλητής. Σύμφωνα με τον Καρακώστα (2002,2004) το λογισμικό μας δίνει κάποιες δυνατότητες οι οποίες είναι οι εξής: «α) No missing values (προεπιλογή). Δεν θεωρείται ελλιπής τιμή καμία παρατήρηση εκτός αυτών με τα κενά κελιά. β) Discrete missing values. Δηλώνονται στα τρία πλαίσια οι 3 διαφορετικές τιμές που όταν καταγράφονται κατά την εισαγωγή των δεδομένων θα σημαίνουν ελλιπή τιμή π.χ. 1, -99, 0. Αυτή η επιλογή είναι πολύ χρήσιμη όταν δεν έχει απαντηθεί μια ερώτηση σε ένα ερωτηματολόγιο είτε γιατί μια συγκεκριμένη ερώτηση δεν διατυπώθηκε στον ερωτώμενο ή ο ερωτώμενος δεν έδωσε απάντηση εσκεμμένα ή κατά λάθος. γ) Range plus one optional discrete missing. Δηλώνεται ένα διάστημα, με κάτω και άνω άκρο τις δηλωθείσες τιμές στα πλαίσια Low και High αντίστοιχα, καθώς και μία διακριτή τιμή. Κάθε τιμή που καταγράφεται εντός του διαστήματος καθώς και η διακριτή τιμή λαμβάνεται ως ελλιπής» (Καρακώστας, 2002; Καρακώστας, 2004).

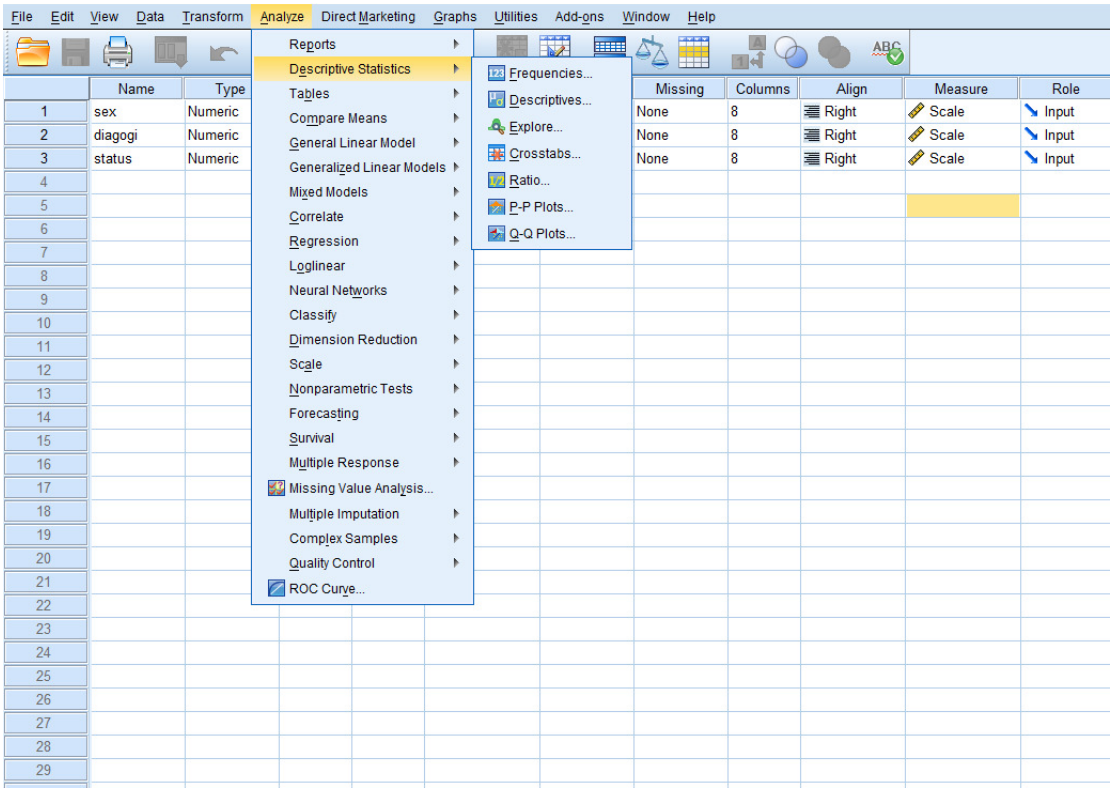
Στη συνέχεια στο κουτί Columns μπορούμε να καθορίσουμε το μήκος κάθε μιας στήλης και στο κουτί Align μπορούμε να στοιχίσουμε τα δεδομένα που θέλουμε μέσα στα κελιά τοποθετώντας τα αριστερά, δεξιά ή στο κέντρο. Επίσης στο κουτί Measure μπορούμε να καθορίσουμε το είδος της μεταβλητής που θα χρησιμοποιήσουμε είτε είναι ποσοτική είτε ποιοτική είτε διατάξιμη, είτε ονοματική. (Χατζηνικολάου, 2002).

Στο σημείο αυτό αξίζει να τονιστεί ότι σημαντικό στάδιο πριν ξεκινήσουμε την ανάλυση της σχέσης μεταξύ δυο ποιοτικών μεταβλητών και γενικότερα σε οποιαδήποτε ανάλυση είναι η αποθήκευση των δεδομένων μας η οποία είναι δυνατό

να επιτευχθεί με τις επιλογές File→Save στη περίπτωση που υπάρχει ήδη το αρχείο δεδομένων και τις επιλογές File→Save as στη περίπτωση που έχουμε ένα νέο αρχείο δεδομένων (Χατζηνικολάου, 2002).

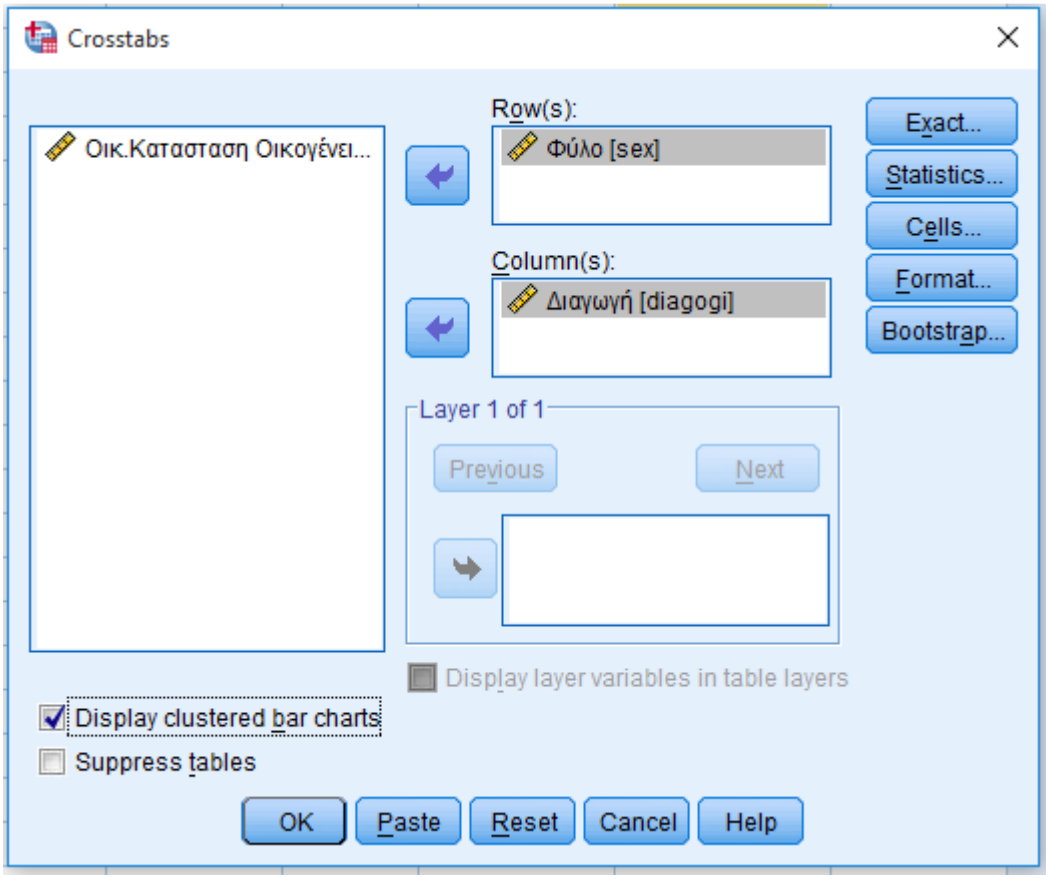
Θέλουμε να εξετάσουμε την ύπαρξη ή όχι σχέσης μεταξύ των ποιοτικών μεταβλητών Φύλο και Διαγωγή. Σε πρώτη φάση αν θέλουμε να εξετάσουμε τα περιγραφικά στατιστικά μέτρα των εξεταζόμενων μεταβλητών για να κατανοήσουμε καλύτερα τα δεδομένα μας ακολουθούμε στο SPSS τα εξής βήματα: Analyze →Descriptive Statistics→ Crosstabs και εμφανίζεται το παρακάτω παράθυρο (Μπατσίδης, 2014).

Εικόνα 5.7: Βήματα για τη περιγραφική στατιστική ανάλυση



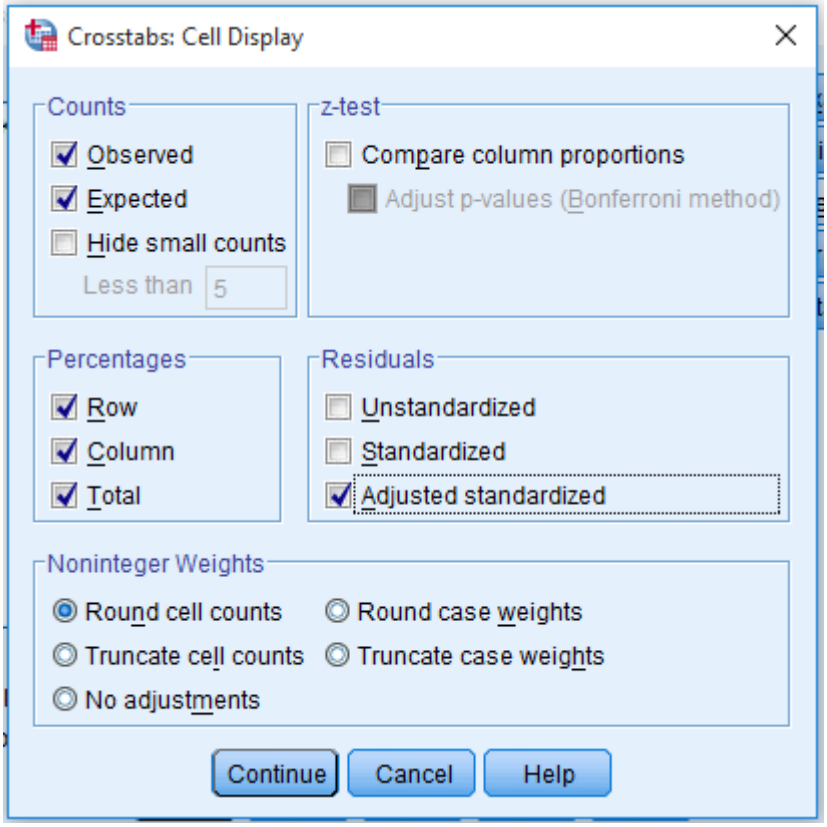
Στη συνέχεια διαλέγουμε την ποιοτική μεταβλητή που μας ενδιαφέρει να εξετάσουμε. Οι τιμές της ποιοτικής μεταβλητής θα βρίσκονται στις γραμμές και στις αντίστοιχες στήλες στο πίνακα συνάφειας. Στη συνέχεια μετακινείται στο πλαίσιο Rows και στο πλαίσιο Columns αντίστοιχα. Κατόπιν έχουμε τη δυνατότητα να κατασκευάσουμε ομάδες ραβδογραμμάτων (bar charts) για κάθε μια τιμή της μεταβλητής η οποία καθορίζεται στο πλαίσιο Rows. Αντίθετα η μεταβλητή που είναι υπεύθυνη για το καθορισμό του ύψους των ράβδων είναι αυτή που έχουμε επιλέξει στο πλαίσιο Columns. Με αυτό τον τρόπο μπορούμε να επιλέξουμε το πλαίσιο Display Cluster Bar Charts φαίνεται στο παρακάτω παράθυρο (Μπατσίδης, 2014).

Εικόνα 5.8: Επιλογή ποιοτικών μεταβλητών (Crosstabs)



Αξίζει να σημειωθεί ότι το πλαίσιο Suppress tables καλό θα ήταν να μη το επιλέγουμε διότι δε θα εμφανιστεί ο πίνακας συνάφειας. Έτσι προχωράει ομαλά η διαδικασία εξέτασης ύπαρξης σχέσης μεταξύ των ποιοτικών μεταβλητών του δείγματος. Όμως για να γίνει αυτό ακόμα πιο σωστά δηλαδή για να αποσαφηνίσουμε την ύπαρξη, την ένταση αλλά και τη φύση της σχέσης μεταξύ των δυο ποιοτικών μεταβλητών είναι σκόπιμο να ενισχύσουμε τις πληροφορίες που μας δίνει το λογισμικό ως προεπιλογή. Αυτό επιτυγχάνεται σε πρώτο στάδιο από την επιλογή Cells επιλέγοντας τα ακόλουθα:

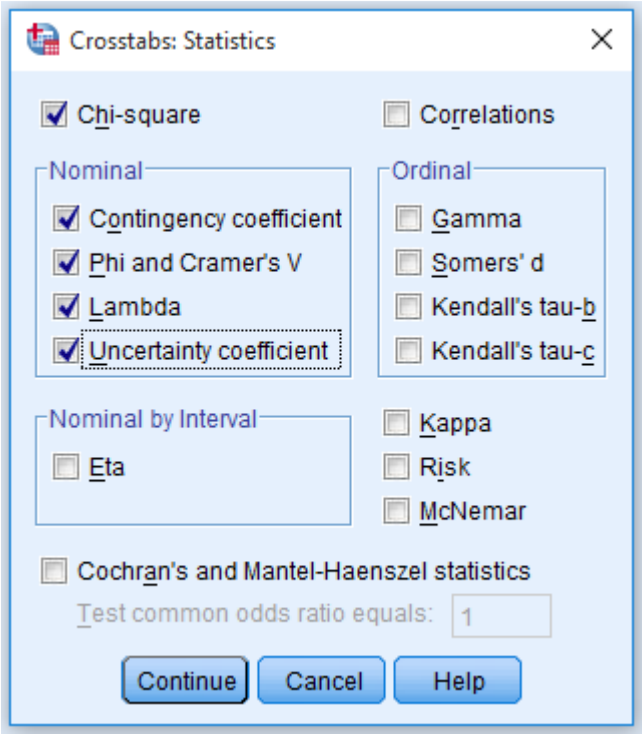
Εικόνα 5.9: Cells



Η επιλογή Observed, Expected counts μας δίνει τη δυνατότητα να αποκτήσουμε τις παρατηρούμενες και αναμενόμενες αντίστοιχα συχνότητες σε κάθε κελί του πίνακα συνάφειας. Τα ποσοστά Percentages μας βοηθούν να αποκτήσουμε τα ποσοστά μέσα στις γραμμές (Row), στις στήλες (Columns) καθώς και για όλα τα δεδομένα (Total). Τα παραπάνω ποσοστά που αντιστοιχίζονται στις γραμμές και στις στήλες έχουν άθροισα 100% και βρίσκονται κατά μήκος των γραμμών και στηλών αντίστοιχα ενώ τα συνολικά ποσοστά έχουν άθροισμα 100% εντός όλων των κελιών του πίνακα.

Επιλέγουμε το πλαίσιο Statistics όπως απεικονίζεται παρακάτω, και μας δίνεται η δυνατότητα να πραγματοποιήσουμε τον έλεγχο ανεξαρτησίας δηλαδή να αναζητήσουμε το βαθμό, τη φύση της συνάφειας καθώς και μια πληθώρα στατιστικών μέτρων. Για ευκολία χρήσης στο παράδειγμα μας επιλέγουμε τα εξής:

Εικόνα 5.10: Στατιστικά μέτρα



Όπως βλέπουμε από το παραπάνω παράθυρο έχουμε επιλέξει όλα τα στατιστικά μέτρα στα οποία περιέχονται συντελεστές συσχέτισης που έχουμε περιγράψει σε προηγούμενες ενότητες.

Αξίζει να αναλυθεί σε αυτό το σημείο όπως επισημαίνει ο Μπατσίδης (2014) ότι: «για πίνακες με 2 γραμμές και 2 στήλες, δηλαδή για ποιοτικές μεταβλητές με δύο δυνατές τιμές η καθεμία, επιλέγοντας το Chi-square όπως φαίνεται παραπάνω μπορούμε να υπολογίσουμε το X^2 του Pearson, το τεστ πηλίκου πιθανοφανειών (the likelihood-ratio chi-square), το Fisher's exact test[7], καθώς και το X^2 τεστ ανεξαρτησίας του Yates με διόρθωση συνεχείας (continuity correction). Για πίνακες συνάφειας μεγαλύτερης διάστασης υπολογίζουμε μόνο το X^2 του Pearson και το τεστ πηλίκου πιθανοφανειών» (Μπατσίδης, 2014).

Επιπρόσθετα το S.P.S.S μας παρέχει τη δυνατότητα να διαπιστώσουμε αν υπάρχουν κελιά που να εμφανίζουν αναμενόμενη τιμή μικρότερη του 5. Σε αυτό το σημείο υπενθυμίζουμε ότι βασική προϋπόθεση προκειμένου να μπορέσουμε να χρησιμοποιήσουμε το X^2, τεστ ανεξαρτησίας του Pearson είναι η μη ύπαρξη αναμενόμενων τιμών που να είναι μικρότερες από το 5. Αλλιώς έχουμε συγχώνευση γειτονικών κελιών, εκτός αν πρόκειται για 2X2 πίνακες όπου καταφεύγουμε στο Fisher's exact test (Μπατσίδης, 2014).

Στη συνέχεια στο παράθυρο στην Εικόνα 5.10 πατάμε continue και είμαστε έτοιμοι να προχωρήσουμε στην εξαγωγή και στην ερμηνεία των αποτελεσμάτων (Πίνακας 5.1).

[7] ένας έλεγχος ιδιαίτερα χρήσιμος για τις περιπτώσεις που δεν ικανοποιούνται οι προϋποθέσεις του X^2 τεστ ανεξαρτησίας

Πίνακας 5.1: Συνάφειας

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Φύλο * Διαγωγή	35	100,0%	0	0,0%	35	100,0%

Από το παραπάνω Πίνακα βλέπουμε ότι ο αριθμός των παρατηρήσεων έχει παραμείνει ο ίδιος (N=35), δεν έχουμε κανένα missing value δηλαδή καμία ελλιπής τιμή. Συνεπώς έχουμε στη διάθεσή μας 35 παρατηρήσεις που ισχύουν ταυτόχρονα και για τις δυο μεταβλητές, Φύλο και Διαγωγή. Ο παραπάνω Πίνακας θεωρείται πίνακας διπλής εισόδου και ονομάζεται Πίνακας Συνάφειας. Στη συνέχεια κάνουμε μια «διασταύρωση» (Crosstabulation) μεταξύ της μεταβλητής Φύλο και της μεταβλητής Διαγωγή. Τα αποτελέσματα φαίνονται στον κάτωθεν πίνακα.

Πίνακας 5.2: Φύλο * Διαγωγή Crosstabulation

Φύλο * Διαγωγή Crosstabulation

			Διαγωγή		Total
			A	B	
Φύλο	Αγόρι	Count	16	3	19
		Expected Count	16,3	2,7	19,0
		% within Φύλο	84,2%	15,8%	100,0%
		% within Διαγωγή	53,3%	60,0%	54,3%
		% of Total	45,7%	8,6%	54,3%
		Adjusted Residual	-,3	,3	
	Κορίτσι	Count	14	2	16
		Expected Count	13,7	2,3	16,0
		% within Φύλο	87,5%	12,5%	100,0%
		% within Διαγωγή	46,7%	40,0%	45,7%
		% of Total	40,0%	5,7%	45,7%
		Adjusted Residual	,3	-,3	
Total		Count	30	5	35
		Expected Count	30,0	5,0	35,0
		% within Φύλο	85,7%	14,3%	100,0%
		% within Διαγωγή	100,0%	100,0%	100,0%
		% of Total	85,7%	14,3%	100,0%

Από τον Πίνακα 5.2 συμπεραίνουμε ότι οι αναμενόμενες συχνότητες (Expected Count) βρίσκονται πολύ κοντά με τις παρατηρούμενες (Count). Επιπλέον στη γραμμή within Φύλο παρατηρούμε ότι το 84,2% των αγοριών στο δείγμα μας παρουσιάζουν διαγωγή Α δηλαδή Κοσμιωτάτη. Αυτό φαίνεται από τη γραμμή within Φύλο και τη διασταύρωση στο πίνακα μεταξύ αγοριού και διαγωγής Α. Επίσης από το πίνακα βλέπουμε ότι το 53,3% που εμφανίζει διαγωγή Κοσμιωτάτη είναι αγόρια. Αυτό φαίνεται από τη γραμμή within Διαγωγή και τη διασταύρωση μεταξύ αγοριού και διαγωγής Α. Επιπλέον αξιοσημείωτο είναι το γεγονός ότι τα αγόρια που παρουσιάζουν διαγωγή Κοσμιωτάτη είναι το 45,7% των συνολικών ερωτηθέντωντου δείγματος. Αυτό αποδεικνύεται από το 45,7% το οποίο βρίσκεται στο συνολικό ποσοστό (Total) και στη διασταύρωση μεταξύ αγοριού και διαγωγής Α. Επιπρόσθετα από το Πίνακα 5.2 παρατηρούμε ότι καμία από τις τιμές των καταλοίπων (Adj. Residuals) δεν είναι μεγαλύτερη σε απόλυτη τιμή από το 1.96.

Στη συνέχεια παρουσιάζεται το Chi-Square Test το οποίο μας δίνει πληροφορίες για το αποτέλεσμα που αφορά τον έλεγχο ανεξαρτησίας.

Πίνακας 5.3: Chi-Square Test

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,077[a]	1	,782		
Continuity Correction[b]	,000	1	1,000		
Likelihood Ratio	,077	1	,781		
Fisher's Exact Test				1,000	,585
Linear-by-Linear Association	,075	1	,785		
N of Valid Cases	35				

a. 2 cells (50,0%) have expected count less than 5. The minimum expected count is 2,29.

b. Computed only for a 2x2 table

Στον Πίνακα 5.3 εξετάζεται η ανεξαρτησία μεταξύ των ποιοτικών μεταβλητών. Βλέπουμε στην υποσημείωση b που δίνεται στο κάτω μέρος του πίνακα ότι υπάρχουν δύο κελιά (50% των συνολικών) με αναμενόμενες συχνότητες μικρότερες του 5. Εφόσον ο πίνακας συνάφειας είναι 2X2 θα χρησιμοποιήσουμε το Fisher's exact test με τη βοήθεια του οποίου διαπιστώνουμε ότι η υπόθεση της ανεξαρτησίας μεταξύ του φύλου και της διαγωγής στο σχολείο δεν απορρίπτεται καθώς η τιμή p-value είναι μεγαλύτερη από 0,05. Στο σημείο αυτό θυμίζουμε ότι όταν η τιμή του συντελεστή είναι μεγαλύτερη από 0,05 δεν απορρίπτουμε τη μηδενική υπόθεση. Αντίθετα όταν η τιμή του συντελεστή είναι μικρότερη από 0,05 απορρίπτουμε τη μηδενική υπόθεση και οι μεταβλητές μας είναι στατιστικά σημαντικές.

Τέλος, στους πίνακες που παρουσιάζονται παρακάτω εμφανίζονται οι τιμές των μέτρων συνάφειας. Όπως είναι αναμενόμενο οι τιμές για αυτούς τους δείκτες είναι κοντά στο μηδέν γεγονός που επιβεβαιώνει ότι η υπόθεση της ανεξαρτησίας δεν απορρίπτεται.

Πίνακας 5.4: Μέτρα συνάφειας-επιβεβαίωση αποτελεσμάτων

			Value	Asymp. Std. Error[a]	Approx. T[d]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,000	,000	.[b]	.[b]
		Φύλο Dependent	,000	,000	.[b]	.[b]
		Διαγωγή Dependent	,000	,000	.[b]	.[b]
	Goodman and Kruskal tau	Φύλο Dependent	,002	,016		,785[c]
		Διαγωγή Dependent	,002	,016		,785[c]
	Uncertainty Coefficient	Symmetric	,002	,014	,140	,781[e]
		Φύλο Dependent	,002	,011	,140	,781[e]
		Διαγωγή Dependent	,003	,019	,140	,781[e]

a. Not assuming the null hypothesis.

b. Cannot be computed because the asymptotic standard error equals zero.

c. Based on chi-square approximation

d. Using the asymptotic standard error assuming the null hypothesis.

e. Likelihood ratio chi-square probability.

Πίνακας 5.5 : Μέτρα συνάφειας

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,047	,782
	Cramer's V	,047	,782
	Contingency Coefficient	,047	,782
N of Valid Cases		35	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

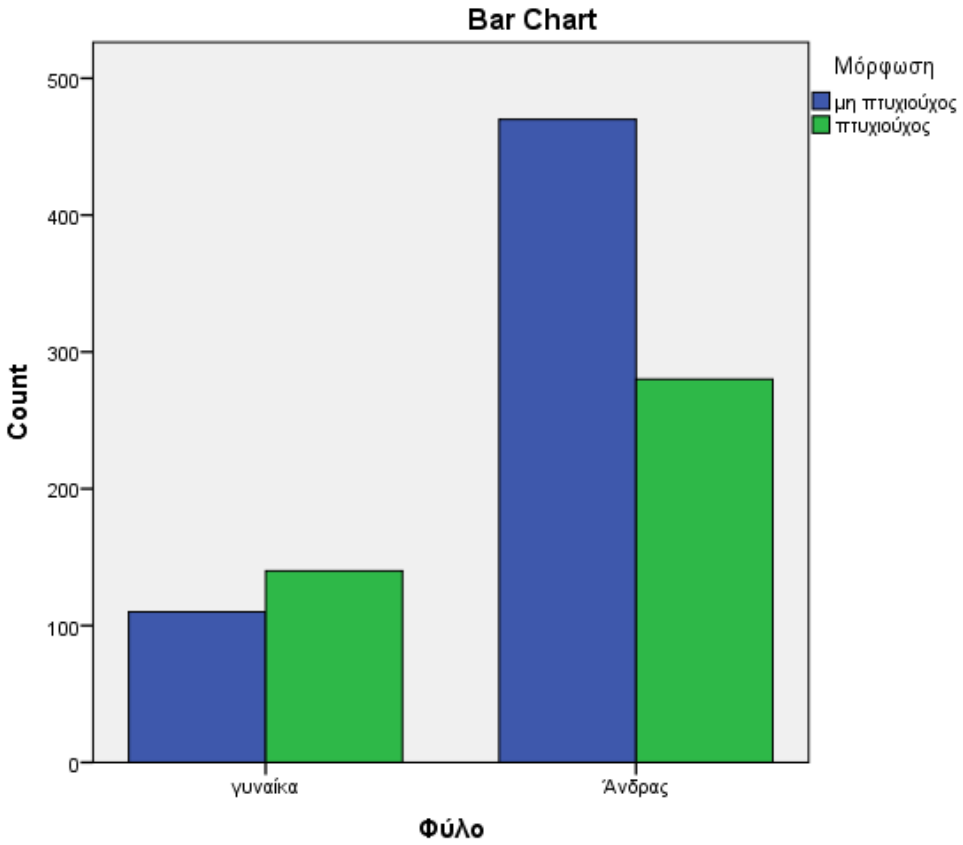
Όπως φαίνεται και από τον Πίνακα 5.5 το p-value παίρνει τη τιμή 0,782 δηλαδή είναι μεγαλύτερη από 0,05 (σε επίπεδο σημαντικότητας 5%). Συνεπώς δεν απορρίπτουμε την μηδενική υπόθεση δηλαδή την υπόθεση ανεξαρτησίας μεταξύ των δυο ποιοτικών μεταβλητών στο δείγμα μας.

Στην περίπτωση που είχαμε ως δεδομένα αυτά που παρουσιάζονται στον ακόλουθο πίνακα διπλής εισόδου και αφορούν το επίπεδο μόρφωσης ανδρών και γυναικών σε επίπεδο πτυχίου, ακολουθείται μια άλλη διαδικασία. Στο Ιστόγραμμα 1 απεικονίζεται η κατανομή των πτυχιούχων και μη πτυχιούχων ανά φύλο.

Πίνακας 5.6: Δεδομένα μη πτυχιούχοι-πτυχιούχοι ανά φύλο

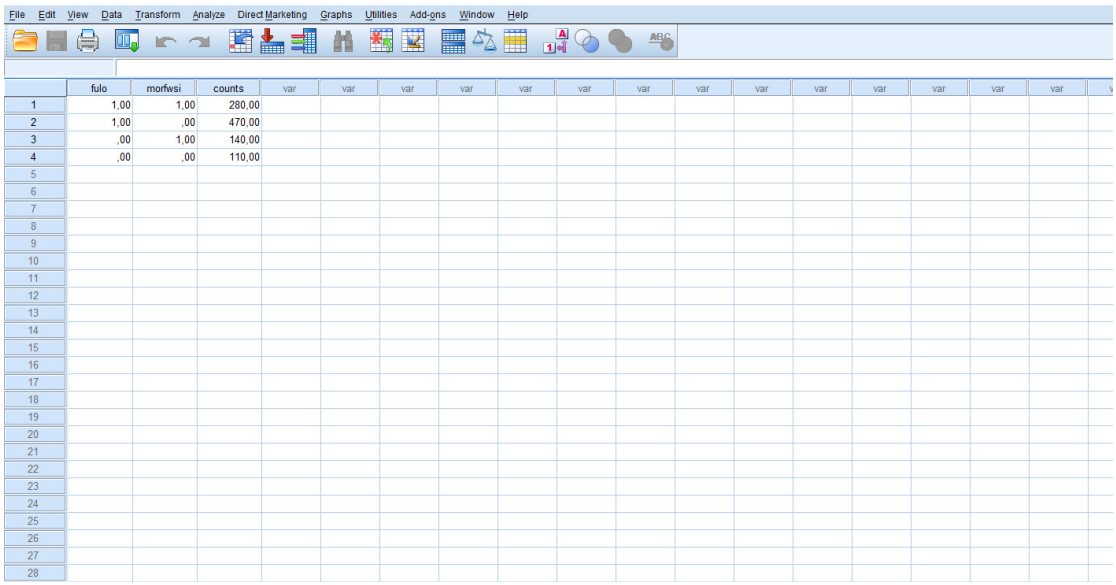
	Μη πτυχιούχοι	Πτυχιούχοι
Άνδρες	470	280
Γυναίκες	110	140

Ιστόγραμμα 1: Πτυχιούχοι-Μη πτυχιούχοι, φύλο



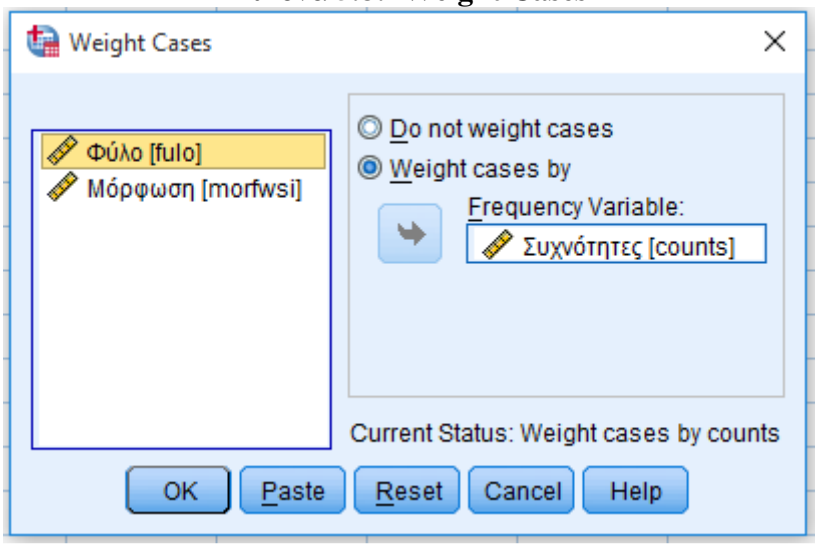
Το ερευνητικό ερώτημα που θέτουμε στην ανάλυσή μας σε αυτό το σημείο είναι αν το φύλο και η κατοχή πτυχίου είναι μεταξύ τους ανεξάρτητα. Θα δούμε παρακάτω τον τρόπο που θα χρησιμοποιηθεί το S.P.S.S. προκειμένου να υπολογιστεί το X^2 τεστ ανεξαρτησίας. Στη προκειμένη περίπτωση στο πλαίσιο με την ένδειξη Data View στις δύο πρώτες στήλες καταγράφονται οι δυνατοί συνδυασμοί των δυο ποιοτικών μεταβλητών που χρησιμοποιούμε. Στο παράδειγμα που εξετάζουμε γίνεται αντιληπτό ότι δημιουργούνται 4 δυνατοί συνδυασμοί καθώς οι τιμές που εξετάζουμε είναι δίτιμες. Συνεπώς για το Φύλο έχουμε 1=άνδρας και 0=γυναίκα, και για τη μεταβλητή Μόρφωση έχουμε 1=πτυχιούχος και 0=μη πτυχιούχος έτσι δημιουργούνται οι δυνατοί συνδυασμοί (1,1), (1,0), (0,1) και (0,0). Στην τρίτη στήλη καταγράφονται οι παρατηρούμενες συχνότητες για κάθε συνδυασμό. Είναι 280, 470, 140 και 110, αντίστοιχα.

Εικόνα 5.7:Καταχώρηση δεδομένων



Σύμφωνα με τον Μπατσίδη (2014) προκειμένου να δηλώσουμε ότι η τρίτη στήλη σε αυτό το σημείο έχει ξεχωριστό ρόλο επιλέγουμε τα εξής βήματα: «Data→ Weight Cases και στο νέο παράθυρο διαλόγου που προκύπτει αφού επιλέξουμε το πλαίσιο Weight cases by τοποθετούμε στο πλαίσιο Frequency Variable τη μεταβλητή όπου καταγράφονται οι παρατηρούμενες συχνότητες και πατάμε OK».

Εικόνα 5.8: Weight Cases



Εν συνεχεία ακολουθούνται τα κλασικά βήματα για να υπολογίσουμε το X^2 τεστ ανεξαρτησίας. Από τα αποτελέσματα που φαίνονται στον επόμενο Πίνακα διαπιστώνουμε ότι το φύλο και η κατοχή πτυχίου δεν είναι μεταξύ τους ανεξάρτητα ενδεχόμενα όπως φαίνεται και από τη P-value του X^2 στατιστικού τεστ η οποία είναι μικρότερη από 0.05 δηλαδή p-value<0.05. Οι γυναίκες που δηλώνουν μη πτυχιούχοι είναι λιγότερες από το αναμενόμενο αποτέλεσμα υπό την ανεξαρτησία (Adj. Residual=-5.2).

Πίνακας 5.9: Φύλο-Μόρφωση (Crosstabulation)

			Μόρφωση		Total
			μη πτυχιούχος	πτυχιούχος	
Φύλο	γυναίκα	Count	110	140	250
		Expected Count	145,0	105,0	250,0
		% within Φύλο	44,0%	56,0%	100,0%
		% within Μόρφωση	19,0%	33,3%	25,0%
		% of Total	11,0%	14,0%	25,0%
		Adjusted Residual	-5,2	5,2	
	Άνδρας	Count	470	280	750
		Expected Count	435,0	315,0	750,0
		% within Φύλο	62,7%	37,3%	100,0%
		% within Μόρφωση	81,0%	66,7%	75,0%
		% of Total	47,0%	28,0%	75,0%
		Adjusted Residual	5,2	-5,2	
Total		Count	580	420	1000
		Expected Count	580,0	420,0	1000,0
		% within Φύλο	58,0%	42,0%	100,0%
		% within Μόρφωση	100,0%	100,0%	100,0%
		% of Total	58,0%	42,0%	100,0%

Στον Πίνακα 5.9 στη γραμμή within Φύλο παρατηρούμε ότι το 44 % των γυναικών δεν είναι πτυχιούχοι (αφού το 44 βρίσκεται στο % within Φύλο και στη διασταύρωση γυναίκας και μη πτυχιούχου). Επίσης στη γραμμή within μόρφωση παρατηρούμε ότι το 19% αυτών που είναι μη πτυχιούχοι είναι γυναίκες (αφού το 19% βρίσκεται στο % within Μόρφωση και στη διασταύρωση γυναίκας και μόρφωσης). Επιπλέον αξιοσημείωτο είναι το γεγονός ότι οι γυναίκες που δεν έχουν πτυχίο αποτελούν το 11% των συνολικών ερωτηθέντων του δείγματος (αφού το 11% βρίσκεται στο % of Total και στη διασταύρωση γυναίκας και μη πτυχιούχος) (Γναρδέλλης, 2003).

Πίνακας 5.10: Chi-Square Tests

Chi-Square Tests

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	26,820[a]	1	,000		
Continuity Correction[b]	26,059	1	,000		
Likelihood Ratio	26,560	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	26,793	1	,000		
N of Valid Cases	1000				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 105,00.

b. Computed only for a 2x2 table

Από τον Πίνακα 5.10 συμπεραίνουμε ότι οι δυο ποιοτικές μεταβλητές που εξετάζουμε δηλαδή το φύλο και το πτυχίο δεν είναι ανεξάρτητες μεταξύ τους. Αυτό επιβεβαιώνεται από το Fisher's Exact Test που η τιμή του είναι 0 συνεπώς απορρίπτουμε τη μηδενική υπόθεση δηλαδή την υπόθεση ανεξαρτησίας. Αφού το p-value είναι μικρότερο από 0.05 απορρίπτουμε τη μηδενική υπόθεση (p-value=0<0.05).

Παρακάτω παραθέτουμε τους πίνακες με τα μέτρα συνάφειας τα οποία επιβεβαιώνουν τα αποτελέσματα μας.

Πίνακας 5.11 : Μέτρα συνάφειας

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,045	,023	1,901	,057
		Φύλο Dependent	,000	,000	.[c]	.[c]
		Μόρφωση Dependent	,071	,036	1,901	,057
	Goodman and Kruskal tau	Φύλο Dependent	,027	,010		,000[d]
		Μόρφωση Dependent	,027	,010		,000[d]
	Uncertainty Coefficient	Symmetric	,021	,008	2,580	,000[e]
		Φύλο Dependent	,024	,009	2,580	,000[e]
		Μόρφωση Dependent	,020	,008	2,580	,000[e]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation

e. Likelihood ratio chi-square probability.

Πίνακας 5.12: Μέτρα συνάφειας

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,164	,000
	Cramer's V	,164	,000
	Contingency Coefficient	,162	,000
N of Valid Cases		1000	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα μελέτη εξετάστηκαν η μέθοδος ανάλυσης της απλής γραμμικής παλινδρόμησης, της πολλαπλής γραμμικής παλινδρόμησης καθώς και ο έλεγχος X^2 ο οποίος εξετάζει τη σχέση μεταξύ ποιοτικών μεταβλητών. Οι παραπάνω μέθοδοι εξετάστηκαν τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο μέσα από την ανάλυση της στατιστικής με τη χρήση του στατιστικού εργαλείου SPSS.

Η μέθοδος της ανάλυσης της απλής γραμμικής παλινδρόμησης αποτελεί μια πιο διαδεδομένη στατιστική τεχνική. Σκοπός της απλής γραμμικής παλινδρόμησης είναι η εξέταση της σχέσης μεταξύ της εξαρτημένης μεταβλητής και της ανεξάρτητης, δηλαδή αποσκοπεί στο να μοντελοποιήσει τη σχέση μεταξύ των δυο μεταβλητών. Μελετάται η επίδραση που ασκεί η ανεξάρτητη μεταβλητή στην εξαρτημένη είτε αυτή είναι θετική είτε αρνητική. Στην πολλαπλή γραμμική παλινδρόμηση χρησιμοποιούνται περισσότερες από μια ανεξάρτητες μεταβλητές. Η πολλαπλή παλινδρόμηση εφαρμόζεται από τους οικονομολόγους με σκοπό να προβλέψει μια μεταβλητή ύστερα από τη μελέτη άλλων μεταβλητών οι οποίες σχετίζονται με γραμμικό τρόπο με αυτή τη μεταβλητή. Έτσι έχουμε την εξίσωση της ευθείας γραμμικής παλινδρόμησης.

Οι μέθοδοι παλινδρόμησης είναι πολύ χρήσιμες για την επίλυση πολλών προβλημάτων που εμφανίζονται στην οικονομία. Αποτελούν ένα πολύ καλό και αξιόπιστο τρόπο πρόβλεψης κάποιων ζητημάτων στην οικονομία μέσα από την εξέταση της σχέσης μεταξύ οικονομικών μεταβλητών. Αυτό δεν είναι πάντα εφικτό διότι τα περισσότερα προβλήματα της φύσης δεν είναι γραμμικά. Η γραμμική προσέγγιση αποτελεί ένα πολύ καλό και αξιόπιστο μοντέλο πρόβλεψης και είναι εύκολο στη χρήση του. Οι τομείς της οικονομίας όπου γίνεται εκτεταμένη χρήση τέτοιου είδους μοντέλων είναι η Μικροοικονομία και Μικροοικονομετρία, η Μακροοικονομία και Μακροοικονομετρία καθώς και η Διαχείριση των Χαρτοφυλακιών (portfolio) και η εξέταση των μετοχών των Χρηματιστηρίων. Για παράδειγμα τα παραπάνω μοντέλα θα μπορούσαν να χρησιμοποιηθούν για να προβλεφθεί το ποσοστό αποταμίευσης μιας χώρας ή το ποσοστό ανεργίας της σε σχέση με το ρυθμό πληθωρισμού της ή το ποσοστό του Ακαθάριστου Εγχώριου Προϊόντος της. Επιπλέον όσον αφορά το τραπεζικό τομέα είναι χρήσιμα για την εξέταση της σχέσης μεταξύ των επιτοκίων της αγοράς και των επενδυτικών δραστηριοτήτων. Οι τραπεζίτες θα ενδιαφέρονταν για την πρόβλεψη των χρηματικών διαθεσίμων στο μέλλον που προέρχονται από τις αποταμιεύσεις. Επίσης τα μοντέλα αυτά συνεισφέρουν σε σημαντικό βαθμό στο εμπόριο μέσα από την εξέταση της επίδρασης της αύξησης των τιμών στη ζήτηση των προϊόντων από τους καταναλωτές. Για τις επιχειρήσεις είναι πολύ χρήσιμη η πληροφορία που λαμβάνουν μέσα από την ανάλυση παλινδρόμησης όσον αφορά τις τιμολογιακές πολιτικές που μπορούν να εφαρμοστούν για τα προϊόντα και τις υπηρεσίες που παρέχουν.

Ο έλεγχος X^2 είναι πολύ σημαντικός για τη διερεύνηση της ύπαρξης σχέσης μεταξύ δυο ποιοτικών μεταβλητών. Ο συγκεκριμένος έλεγχος ελέγχει τα δεδομένα ως προς την καλή προσαρμογή, την ανεξαρτησία και την ομογένειά τους. Είναι σημαντική η χρήση του διότι ελέγχει ως προς την ύπαρξη στατιστικής σημαντικότητας των μεταβλητών που χρησιμοποιεί ένας ερευνητής. Επίσης, ενισχύει την πληροφόρηση του ερευνητή όσον αφορά την ένταση της συσχέτισης μεταξύ των εξεταζόμενων μεταβλητών. Όμως δεν έχει τη δυνατότητα να δώσει ένδειξη για τη κατεύθυνση της συσχέτισης. Συμβάλλει στην εξαγωγή συμπερασμάτων σε έρευνες που διεξάγονται σε διάφορους κλάδους όπως των Οικονομικών, των Επιστημών της Αγωγής κ.α.

Η ανάλυση της παλινδρόμησης δίνει τη δυνατότητα σε όλους τους τομείς της καθημερινότητάς μας να εξαχθούν αποτελέσματα για τη σχέση που συνδέει δύο ή περισσότερες μεταβλητές και οδηγεί σε προβλέψεις που μπορούν να επηρεάσουν το μέλλον της οικονομίας μιας χώρας, μιας επιχείρησης, ενός νοικοκυριού και ούτω καθεξής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Champkin, J., 2013. *Timeline of statistics.,* 21-27.

Gerald, K., 2010 . *"Στατιστική για Οικονομικά & Διοίκηση Επιχειρήσεων".* 8η έκδοση επιμ. Θεσσαλονίκη: Επίκεντρο.

Hendry, D. F., 1980. *Econometrics alchemy or science.* Economica 47,387-406.

Montgomery, D., Peck, E. & Vining, G., 2012. *Introduction to Linear Regression Analysis.* 5th ed. John επιμ.,Wiley & Sons, N. Jersey, 672.

Willcox, W., 1938. *The Founder of Statistics.*. Review of the International Statistical Institute.

Γναρδέλλης, Χ., 2003. *Εφαρμοσμένη Στατιστική.*. Αθήνα: Παπαζήσης.

Ζαχαροπούλου, Χ., 2010. *«Στατιστική – Μέθοδοι και Εφαρμογές», Τόμος Β΄.* Θεσσαλονίκη: Εκδόσεις Σοφία.

Ζωγράφος, Κ. & Γναρδέλλης, Χ., 2003. *Ανάλυση δεδομένων με το PASW Statistics 17.0,* Εκδόσεις Παπαζήση.

Ιωαννίδης, Δ., 2005. *«Στατιστικές Μέθοδοι»,* Θεσσαλονίκη: Εκδόσεις Ζήτη.

Καρακώστας, Κ., 2002. *Μαθήματα Πιθανοτήτων και Στατιστικής.* Πανεπιστήμιο Ιωαννίνων.

Καρακώστας, Κ., 2004. *Γραμμικά Μοντέλα: Παλινδρόμηση, Ανάλυση Διακύμανσης,* Πανεπιστήμιο Ιωαννίνων.

Κικιλίας, Π., Παλαμούρδας, Δ., Πετράκης, Α. & Τσουκαλάς, Δ., 2001. *"Στατιστική - Πιθανότητες".* Αθήνα : Εκδ. Δηρος.

Κιντής, Α., 2010. *Σύγχρονη Οικονομετρική Ανάλυση.* εκδόσεις GUTENBERG επιμ. Αθήνα: Α' ΤΟΜΟΣ.

Μπατσίδης, Α., 2014. *Στατιστική Ανάλυση Δεδομένων : Εισαγωγή και αποθήκευση δεδομένων-Τα βασικά του S.P.S.S.*. Πανεπιστήμιο Ιωαννίνων.

Παπαδόπουλος, Γ., 2008. *Εργαστήριο. Μαθηματικών & Στατιστικής, Ανάλυση Παλινδρόμησης*, (www.aua.gr/gpapadopoulos).

Παπαϊωάννου, Π. & Λουκάς, Σ., 2002. *Εισαγωγή στη Στατιστική.* Εκδόσεις Σταμούλη.

Σιώμκος, Γ. Ι. & Βασιλακοπούλου, Α., 2005. *Εφαρμογή Μεθόδων Ανάλυσης στην Έρευνα Αγοράς.*. Αθήνα: Αθ. Σταμούλης.

Τσαγρής, Μ., 2008. *Στατιστική με τη χρήση του πακέτου SPSS 15.* Αθήνα.

Χάλκος, Γ. Ε., 2011 . *Οικονομετρία, Θεωρία, Εφαρμογές και Χρήση Προγραμμάτων σε Η/Υ*. εκδόσεις GUTENBERG επιμ. Αθήνα.

Χατζηνικολάου, Δ., 2002. *Ανάλυση δεδομένων με τη βοήθεια στατιστικών πακέτων S.P.S.S., Excel, S-Plus.* Εκδόσεις Ζήτη.

Χρήστου, Γ., 2007. *Εισαγωγή στην Οικονομετρία.* Τόμος Β. επιμ. Αθήνα: Εκδόσεις Gutenberg.