Μάθημα 5ο

Παλινδρόμηση

Μέρος Α

Δρ. Αγγελίδης Βασίλης

Στόχοι Ενότητας

- ✓ Απλή Παλινδρόμηση
  - ✓ Εισαγωγή
  - ✓ Προϋποθέσεις - Ορθότητα
- ✓ Two Way Anova (Ανεξάρτητα δείγματα)
- ✓ Χ τετράγωνο

Απλή Παλινδρόμηση - Εισαγωγή

Σε διάφορα προβλήματα της Στατιστικής το ενδιαφέρον μας εστιάζεται στην ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών, για να προσδιορίσουμε με ποιο τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους

Για παράδειγμα

- η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική σχέση μεταξύ τους (*όσο μεγαλώνει η ηλικία του παιδιού μεγαλώνει και το βάρος του*)
- Το ύψος της αμοιβής ενός υπαλλήλου μιας εταιρείας εξαρτάται από το επίπεδο μόρφωσης του, τα χρόνια προϋπηρεσίας, την θέση που κατέχει, κ.λ.π.

Απλή Παλινδρόμηση - Εισαγωγή

Η **ανάλυση παλινδρόμησης** είναι ο κλάδος της Στατιστικής που εξετάζει την σχέση μεταξύ δύο οι περισσότερων μεταβλητών με στόχο την πρόβλεψη της τιμή μιας μεταβλητής από τις τιμές μίας ή πολλών άλλων γνωστών μεταβλητών.

Η μεταβλητή **Y**, που δέχεται την επίδραση της **X** ονομάζεται εξαρτημένη μεταβλητή (*Response variable*). Η μεταβλητή **X** ονομάζεται ανεξάρτητη ή ερμηνευτική μεταβλητή (*Predictor*).

Ο σκοπός της μεθόδου είναι να προσαρμοστούν τα δεδομένα σε ένα υποθετικό προβλεπτικό μοντέλο της σχέσης ανάμεσα στις μεταβλητές. Η γραφική απεικόνιση των τιμών (X_i, Y_i) καλείται διάγραμμα διασποράς (*Scatter plot*).

Απλή Παλινδρόμηση - Εισαγωγή

Απλή ονομάζεται η γραμμική παλινδρόμηση κατά την οποία χρησιμοποιούμε τις τιμές μίας μόνο μεταβλητής (ονομάζεται ερμηνευτική ή προβλεπτική μεταβλητή) για να προβλέψουμε τη μεταβλητή κριτήριο.

Πολλαπλή ονομάζεται η γραμμική παλινδρόμηση κατά την οποία χρησιμοποιούμε τις τιμές πολλών προβλεπτικών μεταβλητών για να προβλέψουμε τη μεταβλητή κριτήριο.

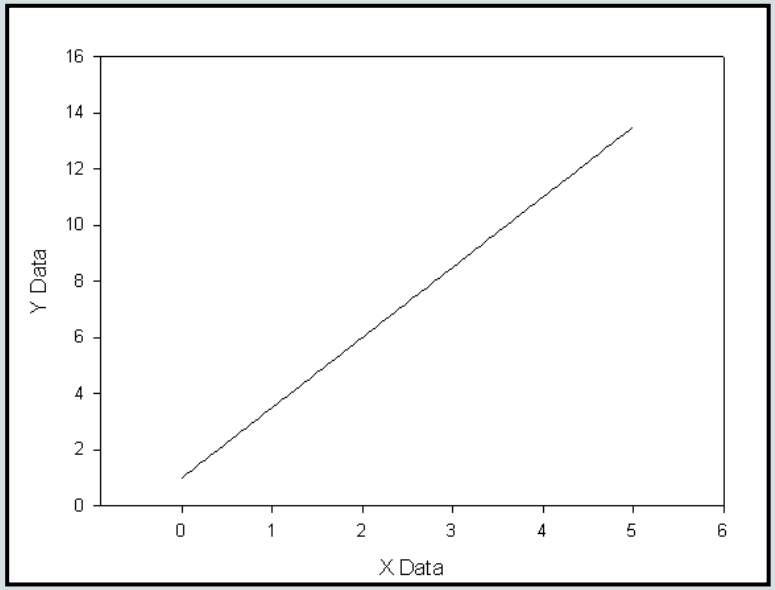
Απλή Παλινδρόμηση - Εισαγωγή

Η σχέση μεταξύ των μεταβλητών **X** και **Y** δεν είναι συναρτησιακή, αλλά στατιστική, δηλαδή οι τιμές της Y δεν ορίζονται μονοσήμαντα από τις αντίστοιχες τιμές της X.

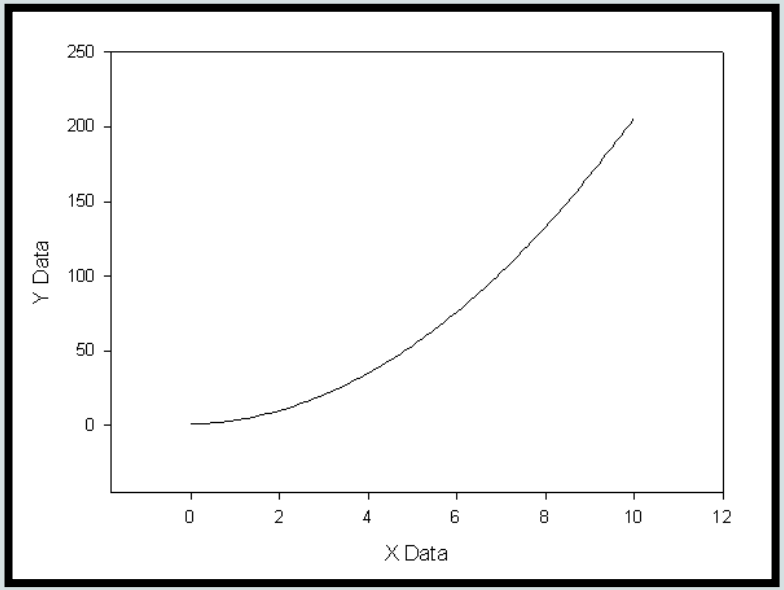
Για παράδειγμα ο μισθός ενός υπαλλήλου σε μια εταιρεία δεν εξαρτάται μόνο από τα χρόνια προϋπηρεσίας του υπάλληλου αλλά και από άλλους παράγοντες όπως το μορφωτικό επίπεδο, την θέση που κατέχει, τυχόν εξειδικεύσεις, οικογενειακή κατάσταση, κ.λ.π.

Στην περίπτωση της γραμμικής παλινδρόμησης, το μοντέλο που εφαρμόζουμε είναι μια ευθεία γραμμή (*Επομένως, περιγράφουμε τη σχέση χρησιμοποιώντας την εξίσωση μιας ευθείας γραμμής*)

Απλή Παλινδρόμηση - Εισαγωγή



Γραμμική Παλινδρόμηση



Μη- Γραμμική Παλινδρόμηση

Απλή Παλινδρόμηση - Εισαγωγή

Αν η σχέση μεταξύ **X** και **Y** είναι **γραμμική**, το υπόδειγμα παλινδρόμησης είναι της μορφής:

$$Y_i = \underbrace{\beta_0 + \beta_1 X}_{\text{όρος παλινδρόμησης}} + \underbrace{\varepsilon_i}_{\text{όρος σφάλματος}}$$

β_0: ο σταθερός όρος (*δηλ. η τιμή του Y όταν το X=0, το σημείο στο οποίο η γραμμή παλινδρόμησης τέμνει τον άξονα Y*)

β_1: ο συντελεστής παλινδρόμησης για την προβλεπτική μεταβλητή ή η κλίση της ευθείας (*δηλ. η γωνία που σχηματίζει η ευθεία με τον άξονα ψ*) ή η κατεύθυνση/δύναμη της σχέσης (*εκφράζει την κατά μέσον όρο μεταβολή της μεταβλητής Y όταν η X μεταβάλλεται κατά μία μονάδα*)

ε_i: Όρος σφάλματος, εκφράζει την απόκλιση των τιμών γύρω από την ευθεία παλινδρόμησης. Ο όρος σφάλματος περιλαμβάνεται, διότι το υπόδειγμα είναι μία προσέγγιση της πραγματικής σχέσης μεταξύ των μεταβλητών.

Απλή Παλινδρόμηση - Εισαγωγή

Έλεγχοι υποθέσεων & Ερμηνεία

Κύριος έλεγχος

H_0: $\beta_1=0$ έναντι της εναλλακτικής **H_1: $\beta_1 \neq 0$**

- ισοδύναμο με τον έλεγχο για συσχέτιση μεταξύ **X** και **Y**
- Δίνει την κλίση της ευθείας Μας ενδιαφέρει για την ερμηνεία των αιτιολογικών σχέσεων μεταξύ φαινομένων – μεταβλητών

Ερμηνεία:

Εξετάζει πόσο αναμένουμε να αυξηθεί η Y με μία μονάδα αύξησης της X

- Η τιμή του β_1 επηρεάζεται από την κλίμακα (μονάδες μέτρησης) των **X** & **Y**.
- Το **ρ** (και r) και ο αντίστοιχος έλεγχος δεν επηρεάζονται

Απλή Παλινδρόμηση - Εισαγωγή

Δευτερεύον Έλεγχος:

H_0: $\beta_0=0$ έναντι της **εναλλακτικής** **H_1: $\beta_0 \neq 0$**

Ερμηνεία:

Η αναμενόμενη τιμή του Y όταν X=0

- Πολλές φορές η τιμή αυτή δεν έχει ερμηνεία (διότι η τιμή $X=0$ δεν παρατηρείται ποτέ στην πράξη). Άλλες φορές θέτουμε $\beta_0=0$ εκ-των-προτέρων και ανεξαρτήτως ελέγχου λόγω κοινής λογικής
- Πολλές φορές «βολεύει» για λόγους ερμηνείας αντί της X να χρησιμοποιήσουμε την στάθμευση.

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

1) **Συνέχεια** των μεταβλητών
   - ✓ Οι μεταβλητές μπορεί να είναι ποσοτικές, είτε διαστήματος (interval) είτε αναλογίας (ratio).
   - ✓ Οι μεταβλητές πρέπει να είναι συνεχείς (Στην περίπτωση κατηγορικών μεταβλητών αυτές εισάγονται στο μοντέλο με μορφή ψευδομεταβλητών)
   - ✓ **Δεν** υπάρχει στατιστικό τεστ για τον έλεγχο βασίζεται στη λογική
2) **Ανεξαρτησία** των παρατηρήσεων
   - ✓ Οι παρατηρήσεις πρέπει να είναι ανεξάρτητες, δηλαδή θα πρέπει να έχει εξασφαλιστεί πως μια παρατήρηση από το ένα δείγμα δεν πρόκειται να ανήκει και στο άλλο
   - ✓ **Δεν** βασίζεται σε κάποιο στατιστικό τεστ, αλλά στη λογική της έρευνας

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

3) Γραμμικότητα των μεταβλητών

- ✓ Πριν την εκτέλεση την γραμμικής παλινδρόμησης πρέπει να γίνει ο έλεγχος της σχέσης μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής
  - Μέσω **διαγράμματος** από το μενού επιλέγουμε Graphs → Legacy Dialogs → Scatter/Dot, και από το παράθυρο που εμφανίζεται την επιλογή 'Simple Scatter' στην συνέχεια βάζουμε την εξαρτημένη μεταβλητή στο κουτί Y Axis και την ανεξάρτητη μεταβλητή στο κουτί X Axis και εκτελούμε την ανάλυση τιμές κοντά στην ευθεία είναι ένδειξη γραμικής σχέσης των μεταβλητών.
  - Μέσω **συντελεστή συσχέτισης** από το μενού επιλέγουμε Analyze → Correlate → Bivariate και στο παράθυρο που εμφανίζεται βάζουμε και τις δύο μεταβλητές στο κουτί variables και τσεκάρουμε το συντελεστή συσχέτισης που θέλουμε να εκτιμήσουμε.

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

4) Κανονικότητα των μεταβλητών

- ✓ Για να ελέγξουμε την κατανομή των μεταβλητών
  - ❖ από το μενού επιλέγουμε Analyze → Descriptive Statistics → P-P Plot (ή Q-Q Plot) κατασκευάζουμε το P-P Plot (ή Q-Q Plot) επιλέγοντας ως μεταβλητή την εξαρτημένη μεταβλητή και ως test distribution την Κανονική κατανομή (Normal) αν τα σημεία βρίσκονται κοντά και εκατέρωθεν της ευθείας δεν υπάρχει γραφική ένδειξη για απόκλιση από την Κανονική κατανομή.
  - ❖ από το μενού επιλέγουμε Analyze → Descriptive Statistics → Explore, και στην συνέχεια από το παράθυρο διαλόγου που εμφανίζεται πατώντας το πλήκτρο Plots επιλέγουμε "*Normality plots with tests*"

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

5) Ανάλυση Καταλοίπων (*Residual Analysis*) → **Κανονικότητα** σφαλμάτων

- ✓ Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο που εμφανίζεται πατώντας το πλήκτρο Save στο κουτί Residuals τσεκάρουμε τις επιλογές Standardized και Studentized, στην συνέχεια εκτελούμε το τεστ Explore για τα κατάλοιπα επιλέγοντας "*Normality plots with tests*" από την από το παράθυρο που εμφανίζεται πατώντας το πλήκτρο Plots
- ✓ Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Plots στο κουτί Standardized Residual Plots τσεκάρουμε τις επιλογές Histogram και Normal Probability Plot
- ✓ Εναλλακτικά μπορούμε και σε αυτή την περίπτωση από το μενού Analyze → Descriptive Statistics → P-P Plot (ή Q-Q Plot) να κατασκευάσουμε το P-P Plot (ή Q-Q Plot) επιλέγοντας ως μεταβλητή τα κατάλοιπα και ως test distribution την Κανονική κατανομή (Normal).

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

6) Ανάλυση Καταλοίπων (*Residual Analysis*) → **Ανεξαρτησία** σφαλμάτων

- ✓ Κατά τη διαδικασία εκτέλεσης του τεστ της απλής παλινδρόμησης (*Linear Regression*) από την επιλογή Save, ζητούμε την αποθήκευση των Standardized Residuals (τυποποιημένα υπόλοιπα), στην συνέχεια από το μενού επιλέγουμε Analyze → Non Parametric Tests→Legacy Dialogs → Runs. Από την p-τιμή του ελέγχου αποφασίζουμε αν υπάρχει αυτοσυσχέτιση ή όχι (*αν p-τιμή>0.05 δεν υπάρχει αυτοσυσχέτιση*).
- ✓ Κατά τη διαδικασία εκτέλεσης του τεστ της απλής παλινδρόμησης (*Linear Regression*) από την επιλογή Statistics τσεκάρουμε στο κουτί Residuals την επιλογή Durbin-Watson. Τιμές κοντά στο 2 υποδηλώνουν άριστη προσαρμογή και κατά συνέπεια μη ύπαρξη αυτοσυσχέτισης. Ενώ τιμές κάτω του 1 και πάνω του τρία υποδηλώνουν θετική και αρνητική αυτοσυσχέτιση αντίστοιχα και δεν πρέπει να συνεχίσουμε την ανάλυση

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

7) Ανάλυση Καταλοίπων (*Residual Analysis*) → **Ομοσκεδαστικότητα** σφαλμάτων

- ✓ Βασική υπόθεση της γραμμικής παλινδρόμησης είναι ότι η διακύμανση των καταλοίπων ε_i παραμένει σταθερή, όποιες και εάν είναι οι τιμές των ερμηνευτικών μεταβλητών.
- ✓ Οπτική διάγνωση της ομοσκεδαστικότητας μπορεί να γίνει κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από την επιλογή Plots, ζητούμε την κατασκευή του διαγραμμάτος διασποράς μεταξύ των Standardized Predicted Values και των Standardized Residuals (μη τυποποιημένες εκτιμώμενες τιμές και μαθητικοποιημένα υπόλοιπα, αντίστοιχα).
- ✓ Αν η διακύμανση είναι σταθερή στο γράφημα που προκύπτει παρατηρούμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από μία οριζόντια γραμμή που περνά από το 0.

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

7) Ανάλυση Καταλοίπων (*Residual Analysis*) → **Ομοσκεδαστικότητα** σφαλμάτων

- ✓ Εναλλακτικά για να δούμε αν παραβιάζεται η ομοσκεδαστικότητα των σφαλμάτων, χωρίζουμε τα σφάλματα σε 2 σχεδόν ισοπληθείς ομάδες (ανάλογα με την τιμή τους) και εξετάζουμε αν οι ομάδες που προέκυψαν παρουσιάζουν στατιστικά ίσες διακυμάνσεις ή όχι.
  - Αρχικά με τη διαδικασία Transform → Recode into different variables δημιουργούμε τη μεταβλητή που κατατάσσει τα Studentized residuals σε δύο ομάδες.
  - Στη συνέχεια κατασκευάζουμε γράφημα διασποράς των Studentized residuals ως προς τις προσαρμοσμένες τιμές, διακρίνοντας με διαφορετικό χρώμα τα studentized residuals που ανήκουν σε διαφορετική ομάδα. Αν η διακύμανση είναι σταθερή στο γράφημα που προκύπτει παρατηρούμε ότι τα υπόλοιπα κατανέμονται τυχαία γύρω από μία οριζόντια γραμμή που περνά από το 0.

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

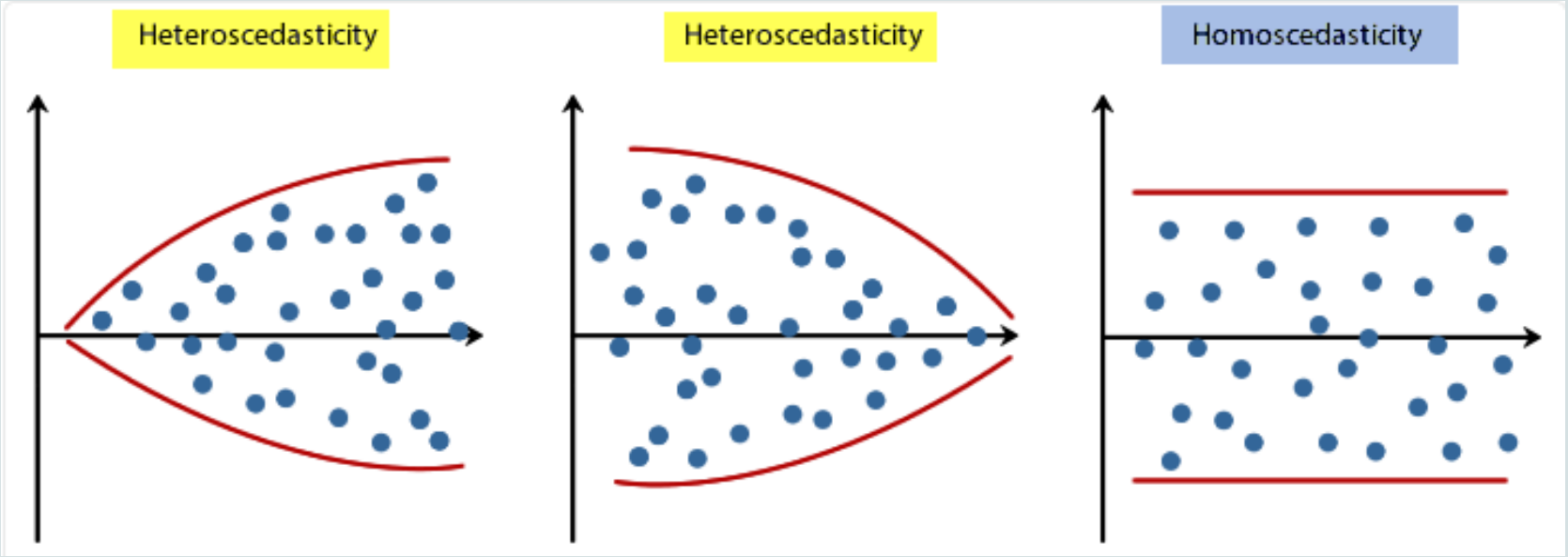
7) Ανάλυση Καταλοίπων (*Residual Analysis*) → **Ομοσκεδαστικότητα** σφαλμάτων

- ✓ Ωστόσο, ο έλεγχος ομοσκεδαστικότητας των σφαλμάτων μπορεί να γίνει και με το τεστ του Levene, που ελέγχει τις ακόλουθες υποθέσεις :
  - ✓ $\mathbf{H_0}$ Τα σφάλματα είναι ομοσκεδαστικά :
  - ✓ $\mathbf{H_1}$ Τα σφάλματα δεν είναι ομοσκεδαστικά.

  Ακολουθώντας τη διαδικασία Analyze → Compare means → Independent samples T-Test χρησιμοποιούμε ως μεταβλητή (test variable) τα Studentized residuals και τα ομαδοποιούμε με τη βοήθεια της μεταβλητής που δημιουργήσαμε παραπάνω.

  Αν το **p − value** είναι μεγαλύτερο του 0,05 δεν υπάρχουν στατιστικές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης της ομοσκεδαστικότητας των σφαλμάτων (σε επίπεδο σημαντικότητας 5%).

7) **Ανάλυση Καταλοίπων** (*Residual Analysis*) → **Ομοσκεδαστικότητα** σφαλμάτων



Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

8) Ακραίες Τιμές (*Outliers*)

- ✓ Ως **ακραία τιμή** χαρακτηρίζουμε μια παρατήρηση η οποία βρίσκεται απροσδόκητα μακριά από τις άλλες παρατηρήσεις. Οι ακραίες παρατηρήσεις μπορούν να ανιχνευθούν αποτελεσματικά με το θηκόγραμμα των παρατηρήσεων ή και με το διάγραμμα υπολοίπων.
- ✓ Όταν στο δείγμα μας περιλαμβάνεται μια τέτοια ακραία τιμή θα πρέπει να ελέγξουμε με προσοχή τα αρχικά δεδομένα για να δούμε αν πράγματι η τιμή αυτή αντιπροσωπεύει ένα πραγματικό σημείο του δείγματος. Αυτό γιατί υπάρχει ενδεχόμενο να έχει γίνει κάποιο λάθος στη διεξαγωγή του πειράματος που να μην είναι δειγματικό λάθος (Για παράδειγμα μπορεί αυτός που καταγράφει τις τιμές να καταγράψει μια τιμή λανθασμένα). Σε μια τέτοια περίπτωση το λάθος θα πρέπει να διορθωθεί ή η τιμή αυτή θα πρέπει να διαγραφεί από τα δεδομένα.

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

8) Ακραίες Τιμές (*Outliers*)

- ✓ Ένας τρόπος ελέγχου της ύπαρξης ή μη ακραίων παρατηρήσεων στα δεδομένα μας γίνεται με τη βοήθεια των τυποποιημένων ή μαθητικοποιημένων υπολοίπων. Τότε παρατηρήσεις των οποίων η απόλυτη τιμή των υπολοίπων αυτών είναι **μεγαλύτερη** του **τρία** (για να είμαστε περισσότερο ακριβείς του 3.29) θεωρούνται ακραίες και συνηθέστερα αποκλείονται από την περαιτέρω ανάλυση. Αν περισσότερο από **1%** των τυποποιημένων υπολοίπων έχουν απόλυτες τιμές μεγαλύτερες του 2.5 (για την ακρίβεια του 2.58) υποδεικνύεται ότι το μοντέλο έχει κακή προσαρμογή. Στο ίδιο συμπέρασμα καταλήγουμε αν **5%** των διαθέσιμων παρατηρήσεων έχουν απόλυτες τιμές των τυποποιημένων υπολοίπων μεγαλύτερες του 2 (του 1.96 για την ακρίβεια όταν το επίπεδο σημαντικότητας είναι 5%).

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

9) Επηρεάζουσες Παρατηρήσεις (*Influential observations*)

- ✓ Είναι πιθανό δύο ή περισσότερες πειραματικές μονάδες να επιδρούν σημαντικά στο μοντέλο παλινδρόμησης. Τέτοιες παρατηρήσεις ονομάζονται **επηρεάζουσες**. Μία τέτοια κατάσταση είναι ανεπιθύμητη καθώς θέλουμε ένα μοντέλο παλινδρόμησης που να μην εξαρτάται από τις τιμές ενός μικρού αριθμού πειραματικών μονάδων, αλλά όλες οι πειραματικές μονάδες να συνεισφέρουν όσο γίνεται το ίδιο στον υπολογισμό των συντελεστών αυτών. Θα πρέπει να δοθεί ξεχωριστή σημασία στις συγκεκριμένες πειραματικές μονάδες που είναι επηρεάζουσες παρατηρήσεις και ίσως πρέπει να παρουσιαστούν τα αποτελέσματα των αναλύσεων με και χωρίς αυτές.

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

9) Επηρεάζουσες Παρατηρήσεις (*Influential observations*)

Από το πλαίσιο Influence Statistics του Linear Regression Save μπορούμε να ζητήσουμε την αποθήκευση διάφορων ποσοτήτων για την εξέταση αυτού του προβλήματος:

- ✓ **DfBeta(s):** Η διαφορά στις τιμές των συντελεστών της παλινδρόμησης αν δεν ληφθεί υπόψη η συγκεκριμένη πειραματική μονάδα. Υπολογίζεται και για τον σταθερό όρο. Οι τυποποιημένες τιμές παρατίθενται στη στήλη Standardized DfBeta. Απόλυτες τιμές αυτών μεγαλύτερες από $\frac{2}{\sqrt{n}}$ μας υποδεικνύουν παρατήρηση που επιδρά στην εκτίμηση των συντελεστών της παλινδρόμησης

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

9) Επηρεάζουσες Παρατηρήσεις (*Influential observations*)

Από το πλαίσιο Influence Statistics του Linear Regression Save μπορούμε να ζητήσουμε την αποθήκευση διάφορων ποσοτήτων για την εξέταση αυτού του προβλήματος:

- ✓ **DfFit:** Μετρά τη διαφορά στην προσαρμογή, δηλαδή στην εκτιμώμενη τιμή, αν δεν συμπεριληφθεί η συγκεκριμένη παρατήρηση στους υπολογισμούς. Δίνονται και οι αντίστοιχες τυποποιημένες τιμές Standardized DfFit. Απόλυτες τιμές αυτών μεγαλύτερες του $2\sqrt{\frac{p+1}{n}}$ υποδεικνύουν επηρεάζουσες παρατηρήσεις

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

9) Επηρεάζουσες Παρατηρήσεις (*Influential observations*)

Από το πλαίσιο Distance του Linear Regression Save μπορούμε να ζητήσουμε την αποθήκευση διάφορων ποσοτήτων για την εξέταση αυτού του προβλήματος:

- ✓ **Leverage:** Τιμές μεγαλύτερες του $\frac{2k+2}{n}$ όπου k είναι ο αριθμός των ανεξάρτητων μεταβλητών και n ο αριθμός των παρατηρήσεων υποδηλώνουν παρατηρήσεις οι οποίες πιθανόν να έχουν μεγάλη επιρροή στην εκτίμηση του μοντέλου και καλύτερα να απομακρυνθούν
- ✓ Cooks: Ένα μέτρο που δείχνει το κατά πόσο τα σφάλματα όλων των παρατηρήσεων θα αλλάξουν αν η συγκεκριμένη παρατήρηση αποκλειστεί από τον υπολογισμό των συντελεστών παλινδρόμησης. Τιμές μεγαλύτερες του $\frac{4}{n-k-1}$ όπου k είναι ο αριθμός των ανεξάρτητων μεταβλητών και n ο αριθμός των παρατηρήσεων υποδηλώνουν επηρεάζουσες παρατηρήσεις

Παλινδρόμηση – Προϋποθέσεις – Ορθότητα

8) Επηρεάζουσες Παρατηρήσεις (*Influences*)

- ✓ Είναι πιθανό δύο ή περισσότερες πειραματικές μονάδες να επιδρούν σημαντικά στο μοντέλο παλινδρόμησης. Τέτοιες παρατηρήσεις ονομάζονται επηρεάζουσες.
- ✓ Εξετάζουμε παρατηρήσεις που έχουν μεγάλη «επιρροή» στο μοντέλο (παρατηρήσεις που αν ληφθούν υπόψη αλλάζουν σημαντικά την εκτίμηση της ευθείας γραμμικής παλινδρόμησης). Τέτοιες παρατηρήσεις είναι αυτές που έχουν Xi αρκετά μακριά από τα υπόλοιπα Xj, $j \neq i$ ή πιο απλά έχουν Xi αρκετά μακριά από το $\bar{X}$. Η «απόσταση» αυτή συνήθως μετράτε χρησιμοποιώντας μια ποσότητα την μόχλευση (leverage) και κυμαίνεται μεταξύ του 0 και του $\frac{n-1}{n}$ όπου n ο αριθμός των παρατηρήσεων. Ο μέσος όρος αυτής της απόστασης είναι $\frac{p}{n}$ όπου p ο αριθμός των ανεξάρτητων μεταβλητών. Επομένως τιμές μεγαλύτερες του $\frac{2p}{n}$ πρέπει να απορριφθούν

Απλή Παλινδρόμηση – Παράδειγμα Ι

Ο υπεύθυνος του γραφείου εξυπηρέτησης πελατών μιας εταιρείας, ενδιαφέρεται να εκτιμήσει το χρόνο που μεσολαβεί από την παραγγελία έως την παράδοση (άρα και το αντίστοιχο κόστος αλλά και την ποιότητα εξυπηρέτησης) κάθε παραγγελίας ανάλογα με την απόσταση του πελάτη από τις κεντρικές αποθήκες της εταιρείας. Για το λόγο αυτό πήρε ένα τυχαίο δείγμα 10 παραγγελιών και κατέγραψε την απόσταση των εγκαταστάσεων του πελάτη (σε χιλιόμετρα) και τις ημέρες που μεσολάβησαν μέχρι την παράδοση των εμπορευμάτων παράδοσης. Να κατασκευαστεί ένα μοντέλο που θα βοηθήσει τον υπεύθυνο της εταιρείας (αρχείο δεδομένων paradigma_1.sav)

Απόσταση	825	215	1070	550	480	920	1350	325	670	1215
Ημέρες	3,5	1	4	2	1,5	4	4,5	1,5	3	5,5
Απόσταση	770	280	1060	710	120	200	350	875	730	600
Ημέρες	3	1	5	3	0,5	1	1,6	3,6	3	2,8

Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα

Το παράδειγμα περιέχει δύο μεταβλητές ποσοτικές την απόσταση (ανεξάρτητη) και τον χρόνο παράδοσης (εξαρτημένη). Αρχικά ελέγχουμε αν ένα μοντέλο απλής παλινδρόμησης είναι κατάλληλο για την συγκεκριμένη περίπτωση ελέγχοντας τις προϋποθέσεις ορθότητας χρήσης του μοντέλου

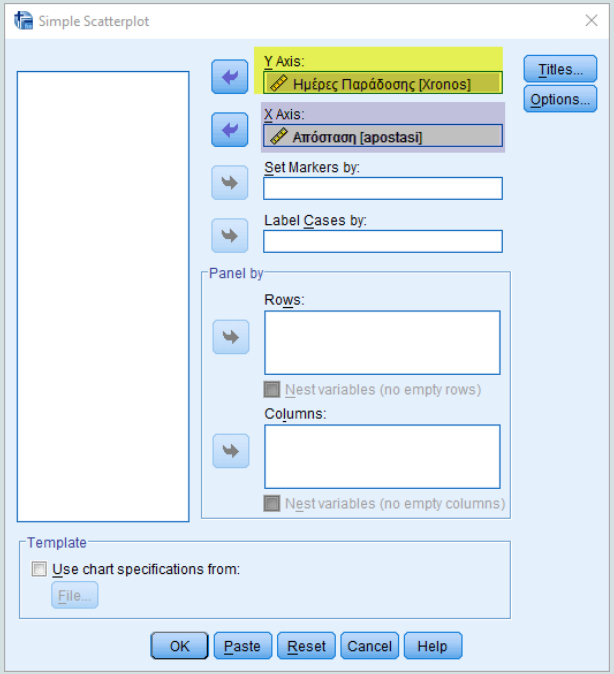
1) Η **Συνέχεια** των μεταβλητών προκύπτει από τα δεδομένα
2) Η **Ανεξαρτησία** των παρατηρήσεων προκύπτει από την περιγραφή του παραδείγματος
3) Για την **Γραμμικότητα** μεταξύ των δύο μεταβλητών
    i. Θα κατασκευάσουμε ένα **γράφημα** διασποράς (Scatter plot) της εξαρτημένης με την ανεξάρτητη
    ii. Θα υπολογίσουμε τον **συντελεστή** συσχέτισης

Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – **Γραμμικότητα** (3)

Γράφημα διασποράς (Scatter plot) της εξαρτημένης με την ανεξάρτητη

Από το μενού επιλέγουμε Graphs → Legacy Dialogs → Scatter/Dot, και από το παράθυρο που εμφανίζεται την επιλογή 'Simple Scatter' στην συνέχεια βάζουμε την εξαρτημένη μεταβλητή (xronos) στο κουτί Y Axis και την ανεξάρτητη μεταβλητή (apostasy) στο κουτί X Axis και εκτελούμε την ανάλυση

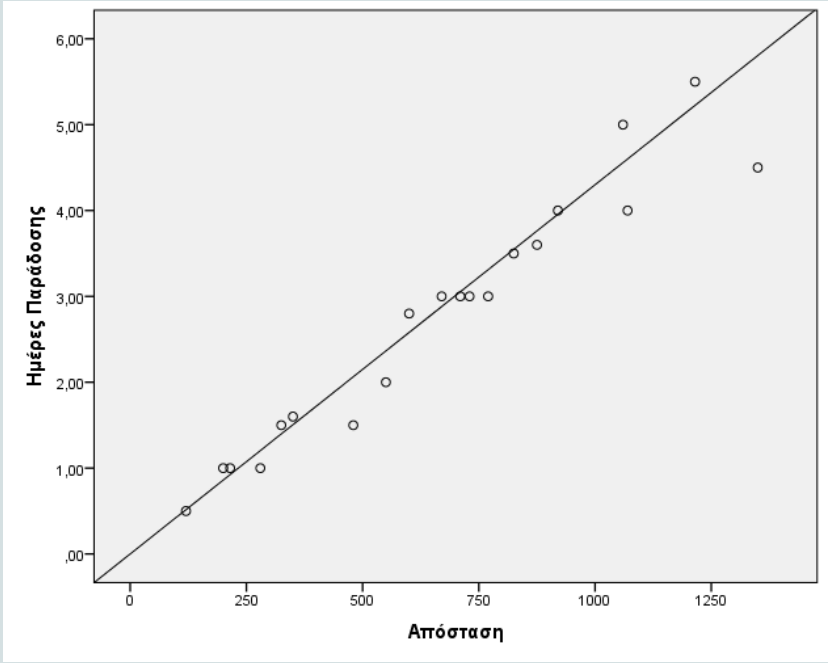


Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – **Γραμμικότητα** (3)

Γράφημα διασποράς (Scatter plot) της εξαρτημένης με την ανεξάρτητη

Από γράφημα διασποράς παρατηρούμε ότι όλες οι τιμές βρίσκονται σε κοντινή απόσταση εκατέρωθεν της ευθείας που είναι μια ένδειξη ύπαρξης γραμμικής σχέσης μεταξύ των δύο μεταβλητών.

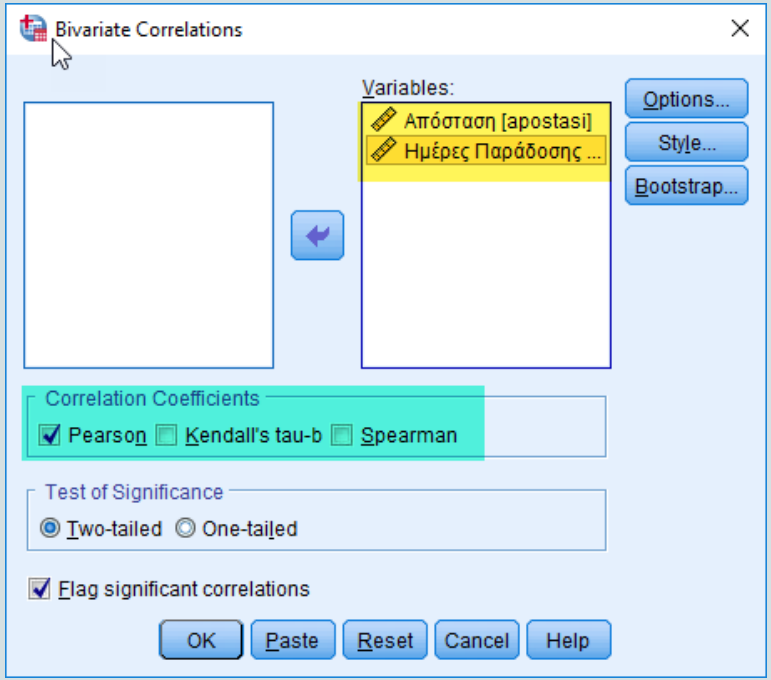


Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – Γραμμικότητα (3)

Συντελεστής συσχέτισης

Από το κεντρικό μενού επιλέγουμε Analyze → Correlate → Bivariate και στο παράθυρο που εμφανίζεται βάζουμε και τις δύο μεταβλητές στο κουτί variables (apostasi, xronos) και τσεκάρουμε το συντελεστή συσχέτισης που θέλουμε να εκτιμήσουμε (Pearson) .



Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Γραμμικότητα** (3)

Συντελεστής συσχέτισης

Από τα αποτελέσματα παρατηρούμε ότι ο συντελεστής συσχέτισης (**r=0,967**) είναι πολύ υψηλός (όσο πιο μεγάλος ο δείκτης αυτός, τόσο ισχυρότερη είναι η συσχέτιση των δύο μεταβλητών (θετική ή αρνητική)

Correlations

		Απόσταση	Ημέρες Παράδοσης
Απόσταση	Pearson Correlation	1	,967**
	Sig. (2-tailed)		,000
	N	20	20
Ημέρες Παράδοσης	Pearson Correlation	,967**	1
	Sig. (2-tailed)	,000	
	N	20	20

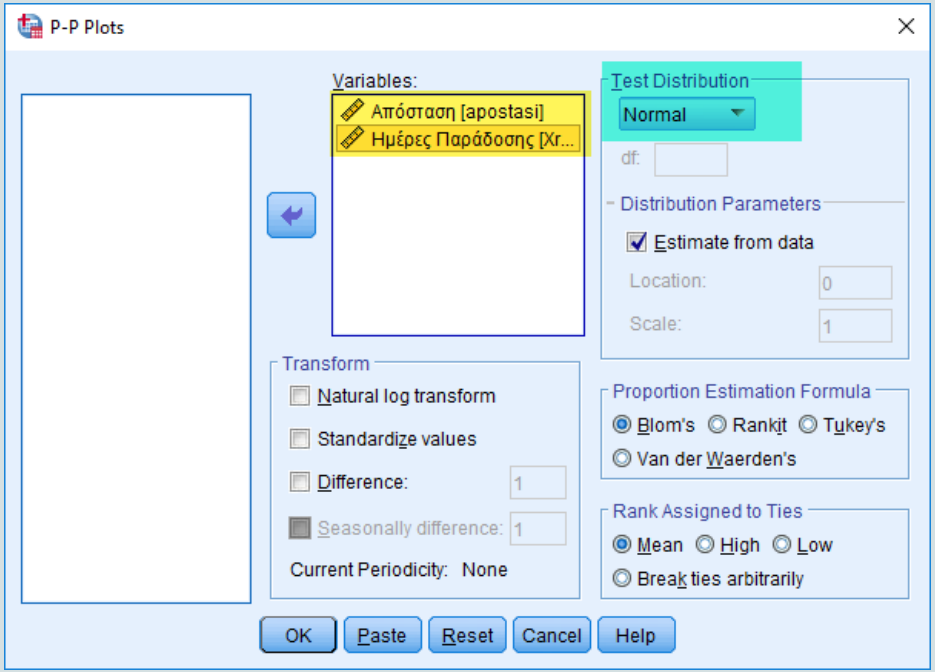
**. Correlation is significant at the 0.01 level (2-tailed).

Η σχέση επίσης είναι στατιστικά σημαντική απορρίπτοντας την μηδενική $H_0: \rho = 0$, δηλαδή παρατηρείται μια ισχυρή (θετική) γραμμική συσχέτιση μεταξύ των δύο μεταβλητών

Απλή Παλινδρόμηση – Παράδειγμα Ι

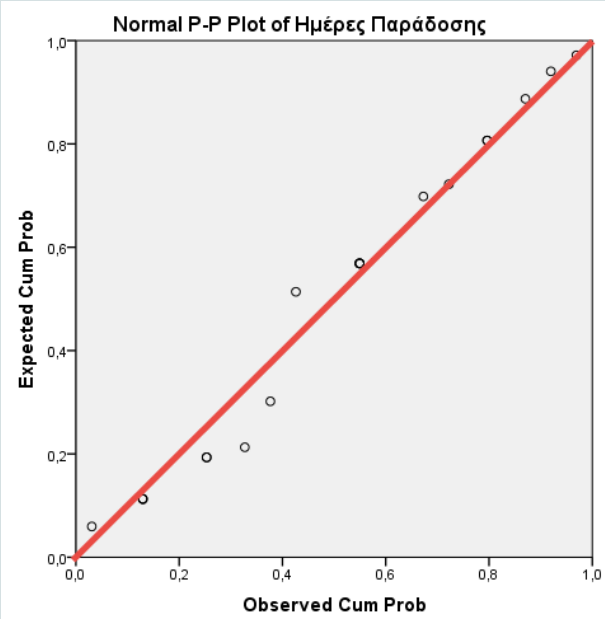
Προϋποθέσεις – Ορθότητα – **Κανονικότητα μεταβλητών** (4)

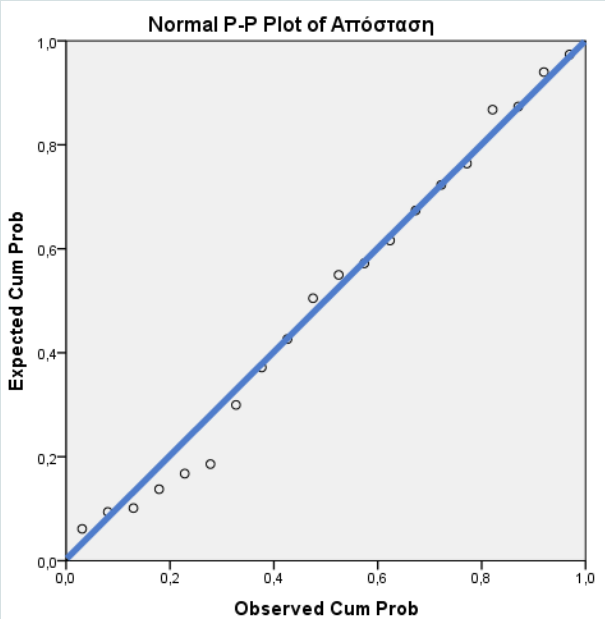
Από το μενού επιλέγουμε Analyze → Descriptive Statistics → P-P Plot (ή Q-Q Plot) κατασκευάζουμε το P-P Plot (ή Q-Q Plot) για την εξαρτημένη και την ανεξάρτητη μεταβλητή και ως test distribution την Κανονική κατανομή (Normal)



Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Κανονικότητα μεταβλητών** (4)



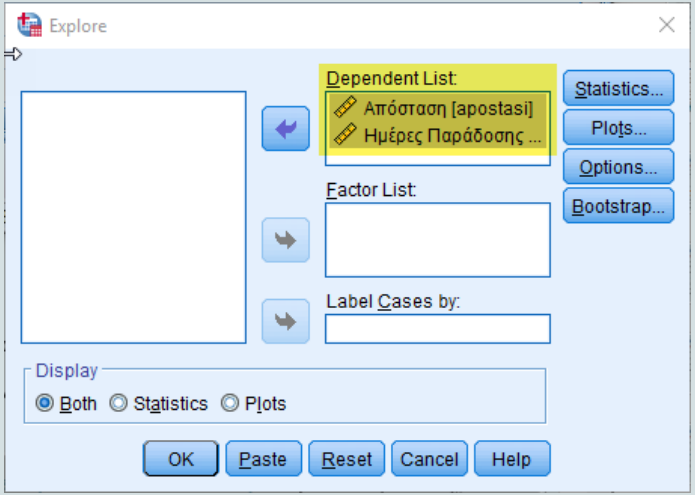


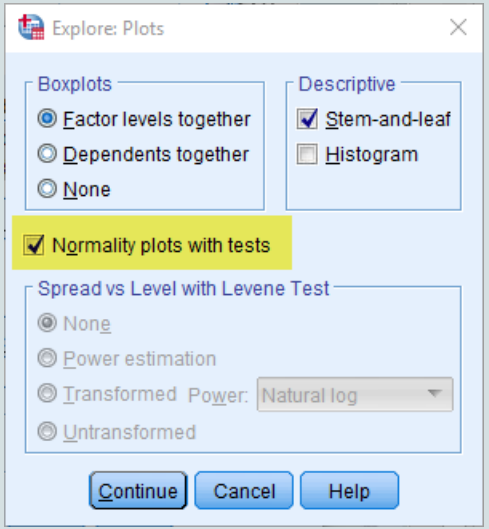
Από τα διαγράμματα διασποράς τα σημεία βρίσκονται κοντά και εκατέρωθεν της ευθείας οπότε **δεν** υπάρχει γραφική ένδειξη για απόκλιση από την Κανονική κατανομή.

Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – **Κανονικότητα μεταβλητών** (4)

Από το μενού επιλέγουμε Analyze → Descriptive Statistics → Explore στο πλαίσιο Dependent List προσθέτουμε τις δύο μεταβλητές





στην συνέχεια από το παράθυρο διαλόγου που εμφανίζεται πατώντας το πλήκτρο Plots επιλέγουμε "*Normality plots with tests*"

Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Κανονικότητα μεταβλητών** (4)

Μεγάλα Δείγματα

Μικρά Δείγματα

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Απόσταση	,114	20	,200*	,969	20	,742
Ημέρες Παράδοσης	,137	20	,200*	,954	20	,437

*. This is a lower bound of the true significance.

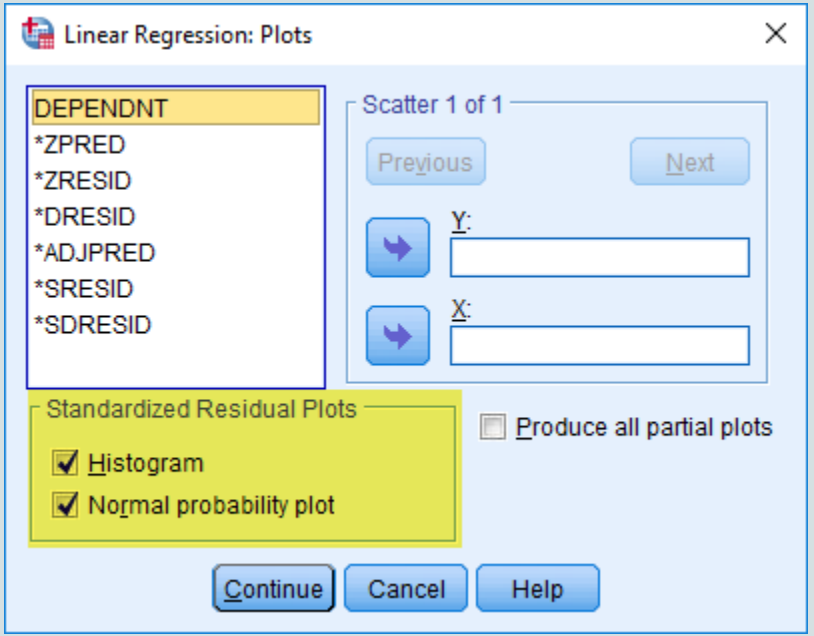
a. Lilliefors Significance Correction

Από το αποτελέσματα της ανάλυσης παρατηρούμε ότι τα αντίστοιχα **p values** είναι μεγαλύτερα της στατιστικής σημαντικότητας (0,05) οπότε **δεν** απορρίπτουμε την υπόθεση της ακολουθίας κανονικής κατανομής για καμία από τις δύο μεταβλητές

Απλή Παλινδρόμηση – Παράδειγμα Ι

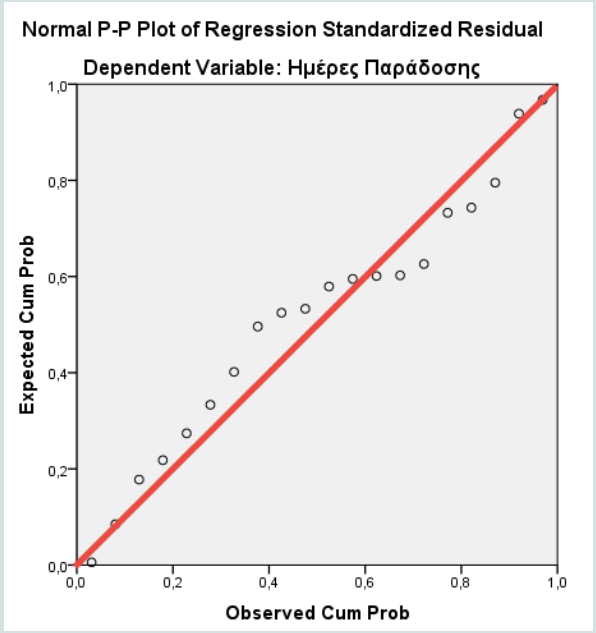
Προϋποθέσεις – Ορθότητα – **Κανονικότητα Σφαλμάτων** (5)

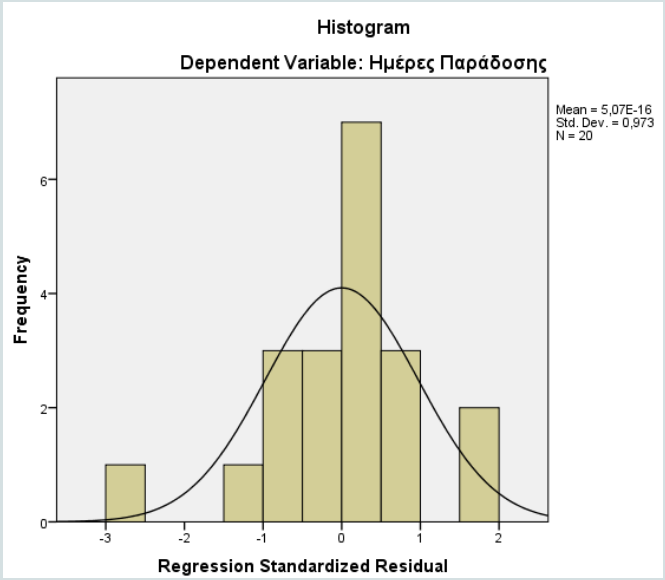
Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο Plots στο κουτί Standardized Residual Plots τσεκάρουμε τις επιλογές Histogram και Normal Probability Plot



Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Κανονικότητα Σφαλμάτων** (5)





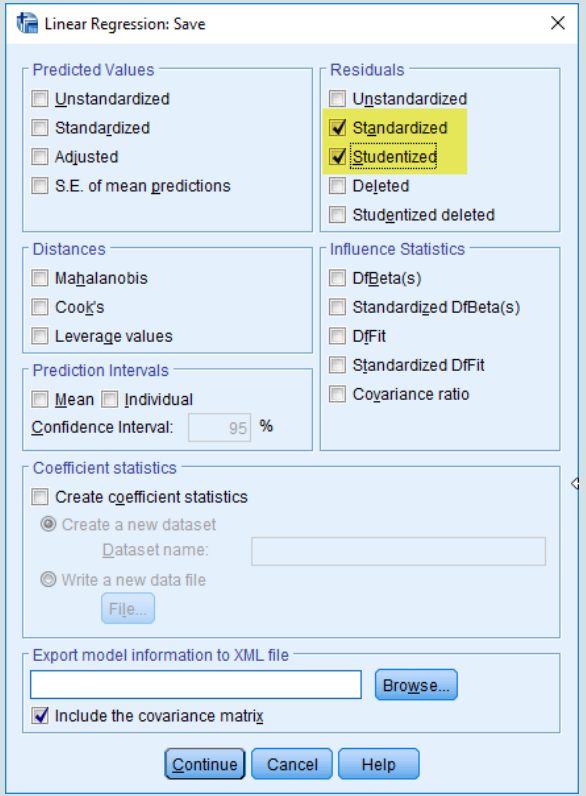
Από τα αποτελέσματα παρατηρούμε ότι **δεν** υπάρχουν ενδείξεις για να απορρίψουμε την κανονική κατανομή των καταλοίπων

Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Κανονικότητα Σφαλμάτων** (5)

Κατά τη διαδικασία προσαρμογής του μοντέλου της παλινδρόμησης από το παράθυρο που εμφανίζεται πατώντας το πλήκτρο Save στο κουτί Residuals τσεκάρουμε τις επιλογές Standardized και Studentized, και εκτελούμε την ανάλυση

Στην συνέχεια εκτελούμε το τεστ Explore για τα κατάλοιπα επιλέγοντας "*Normality plots with tests*" από την από το παράθυρο που εμφανίζεται πατώντας το πλήκτρο Plots



Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Κανονικότητα Σφαλμάτων** (5)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,146	20	,200*	,953	20	,421
Studentized Residual	,149	20	,200*	,938	20	,220

*. This is a lower bound of the true significance.

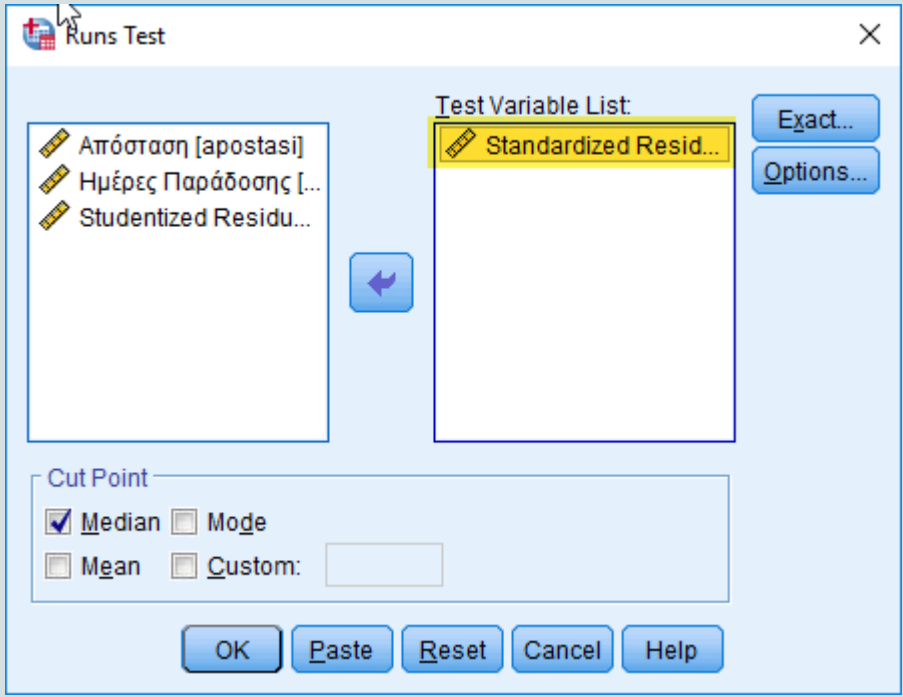
a. Lilliefors Significance Correction

Από το αποτελέσματα της ανάλυσης παρατηρούμε ότι τα αντίστοιχα **p values** είναι μεγαλύτερα της στατιστικής σημαντικότητας (0,05) οπότε **δεν** απορρίπτουμε την μηδενική και αποδεχόμαστε την υπόθεση της κανονικότητας για τα κατάλοιπα

Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – Ανεξαρτησία Σφαλμάτων (6)

Κατά τη διαδικασία εκτέλεσης του τεστ της απλής παλινδρόμησης (*Linear Regression*) από την επιλογή Save, ζητούμε την αποθήκευση των Standardized Residuals (τυποποιημένα υπόλοιπα), στην συνέχεια από το κεντρικό μενού επιλέγουμε Analyze → Non Parametric Tests→Legacy Dialogs → Runs.



Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – **Ανεξαρτησία Σφαλμάτων** (6)

Με βάση το παραπάνω τεστ το οποίο εμφανίζεται στο διπλανό πίνακα κα έχει **p-value = 0,094**> 0,05 οπότε δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση δηλαδή ότι κατάλοιπα είναι τυχαία.

Runs Test

	Standardized Residual
Test Value[a]	,14125
Cases < Test Value	10
Cases >= Test Value	10
Total Cases	20
Number of Runs	11
Z	,000
Asymp. Sig. (2-tailed)	1,000

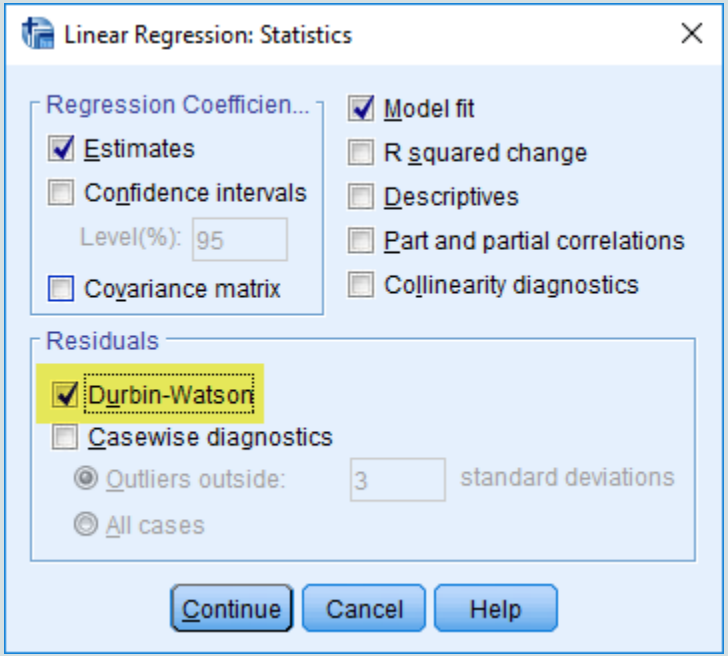
a. Median

Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – **Ανεξαρτησία Σφαλμάτων** (6)

Κατά τη διαδικασία εκτέλεσης του τεστ της απλής παλινδρόμησης (*Linear Regression*) από την επιλογή Statistics τσεκάρουμε στο κουτί Residuals την επιλογή Durbin-Watson.

Ο δείκτης εμφανίζεται στον πίνακα model summary



Απλή Παλινδρόμηση – Παράδειγμα I

Προϋποθέσεις – Ορθότητα – **Ανεξαρτησία Σφαλμάτων** (6)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,967[a]	,936	,932	,37584	2,127

a. Predictors: (Constant), Απόσταση

b. Dependent Variable: Ημέρες Παράδοσης

Η τιμή του δείκτη **Durbin Watson** όπως φαίνεται στον διπλανό πίνακα είναι **2,127** – τιμή η οποία υποδεικνύει μια άριστη προσαρμογή των δεδομένων και ανεξαρτησία των καταλοίπων

Απλή Παλινδρόμηση – Παράδειγμα Ι

Προϋποθέσεις – Ορθότητα – **Ομοσκεδαστικότητα Σφαλμάτων** (7)

Κατά τη διαδικασία εκτέλεσης του τεστ της απλής παλινδρόμησης (*Linear Regression*) από την επιλογή Plots εισάγουμε θέση Y: τα standardized residual (ZRESIG) και στην θέση X: τα standarixed τσεκάρουμε στο κουτί Residuals την επιλογή Durbin-Watson.

Ο δείκτης εμφανίζεται στον πίνακα model summary