Πανεπιστήμιο Θεσσαλίας

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Χωροταξίας, Πολεοδομίας & Περιφερειακής Ανάπτυξης

ΜΑΘΗΜΑ ΕΠΙΛΟΓΗΣ: ΟΙΚΟΝΟΜΕΤΡΙΑ

Το Γενικευμένο Γραμμικό Υπόδειγμα (Α)

ΔΙΑΛΕΞΗ 05

Μαρί-Νοέλ Ντυκέν, Μαρία Τσιάπα
mdyken@prd.uth.gr, mtsiapa@prd.uth.gr

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Τι επιτρέπει;

- ✓ Να θέσουμε υπό έλεγχο πολλούς παράγοντες που *ceteris paribus* αναμένονται να επηρεάζουν ταυτόχρονα το φαινόμενο που εξετάζουμε (εξαρτημένη μεταβλητή).

- ✓ Ενσωματώνοντας πολλές ερμηνευτικές μεταβλητές, θα είμαστε σε θέση να ερμηνεύσουμε μεγαλύτερο ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής.

- ✓ Όμως ενδέχεται οι ερμηνευτικές μεταβλητές (η τουλάχιστον ορισμένες από αυτές) να συσχετίζονται μεταξύ τους με αποτέλεσμα, η ανάλυση και οι εκτιμήσεις να είναι παραπλανητικές.

Γενίκευση του «Γραμμικού Υποδείγματος»

ΣΤΟΧΟΣ

- ✓ Εξειδίκευση του υποδείγματος: Μαθηματική μορφή
- ✓ Αξιολόγηση του πολυμεταβλητού υποδείγματος
- ✓ Το ζήτημα της «***ανεξαρτησίας***» των ερμηνευτικών μεταβλητών
- ✓ Αναζήτηση «Πολυσυγγραμικότητας»
- ✓ Εφαρμογή στο SPSS: ερμηνεία των αποτελεσμάτων

Εξειδίκευση του υποδείγματος

Μαθηματική μορφή

[1] $\mathbf{Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_j X_{ij} + \ldots + b_k X_{ik} + \varepsilon_i \quad i = 1, \ldots, n \quad j = 1, \ldots, k}$

n = αριθμός παρατηρήσεων, k = αριθμός ερμηνευτικών μεταβλητών

b_j = Ξεχωριστή επίδραση των ανεξαρτήτων μεταβλητών όταν οι άλλες είναι σταθερές (ceteris paribus)

$b_j = dY / d\,\mathbf{X_j}$

Υπό μορφή Μήτρων

[2] $\mathbf{Y = X.b + \varepsilon}$

Y = Διάνυσμα στήλης (n,1)

X = Μήτρα (n,k)

b = Διάνυσμα στήλης (n,1)

ε = Διάνυσμα στήλης (n,1)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ .. \\ Y_i \\ .. \\ .. \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{10} & X_{11} & X_{1j} & X_{1k} \\ X_{20} & X_{21} & X_{2j} & X_{2k} \\ .. & .. & .. & .. \\ .. & .. & .. & .. \\ X_{i0} & X_{i1} & X_{ij} & X_{ik} \\ .. & .. & .. & .. \\ .. & .. & .. & .. \\ X_{n0} & X_{n1} & X_{nj} & X_{nk} \end{bmatrix} x \begin{bmatrix} b_0 \\ b_1 \\ .. \\ .. \\ b_i \\ .. \\ .. \\ b_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ .. \\ .. \\ \varepsilon_i \\ .. \\ .. \\ \varepsilon_k \end{bmatrix}$$

Οι Υποθέσεις του υποδείγματος

Γραμμικότητα

Από τις σχέσεις [1] και [2], προκύπτει ότι, η εξαρτημένη μεταβλητή είναι γραμμική συνάρτηση των ανεξαρτήτων μεταβλητών .

Διαταρακτικός όρος

Ο μέσος του διαταρακτικού όρου = 0

Η διακύμανση του διαταρακτικού όρου = σ^2, σταθερή

Η συνδιακύμανση των διαδοχικών τιμών του διαταρακτικού όρου = 0

- $V[\varepsilon_i] = \sigma^2$, Δεν υπάρχει Ετεροσκεδαστικότητα
- $COV[\varepsilon_i, \varepsilon_j] = 0$, Δεν υπάρχει Αυτοσυσχέτιση(*)

Κατά συνέπεια:

$\varepsilon_i \Longrightarrow N(0, \sigma^2)$

Η ετεροσκεδαστικότητα όπως και η αυτοσυσχέτιση είναι σοβαρά προβλήματα επειδή η μέθοδος των ελαχίστων τετραγώνων (OLS) υποθέτει ότι όλα τα κατάλοιπα αντλούνται από έναν πληθυσμό που έχει μια σταθερή διακύμανση (homoscedasticity) και τα κατάλοιπα δεν συσχετίζονται μεταξύ τους.

Η ετεροσκεδαστικότητα αποτελεί συχνό φαινόμενο όπως και η αυτοσυσχέτιση η οποία εμφανίζεται σχεδόν συστηματικά με χρονολογικές σειρές

Οι Υποθέσεις του υποδείγματος

Ερμηνευτικές μεταβλητές

Οι ερμηνευτικές μεταβλητές (η μήτρα Χ) δεν είναι στοχαστικές δηλαδή οι τιμές της κάθε μιας μεταβλητής παραμένουν σταθερές (άλλα όχι ίσες μεταξύ τους) σε επαναλαμβανόμενα δείγματα.
Αυτό σημαίνει ότι, αν έχουμε διάφορα δείγματα (ίδιο μέγεθος = n) για την εξαρτημένη Υ και τη μήτρα Χ, θεωρούμε οι τιμές της μήτρας δεν μεταβάλλονται από δείγμα σε δείγμα.

Δεν υπάρχει ακριβής γραμμική σχέση ανάμεσα στις k ανεξάρτητες μεταβλητές $\mathbf{X_j}$. Πρόκειται για μια από τις σοβαρότατες υποθέσεις της παλινδρόμησης.

Η υπόθεση αυτή αναφέρεται ως απουσία πλήρους πολυσυγγραμμικότητας (multicollinearity) η οποία θα πρέπει να ελεγχθεί συστηματικά.

Συνηθώς, η πολυσυγγραμμικότητα αποτελεί το πρώτο θέμα που εξετάζουμε όταν χρησιμοποιούμε την γραμμική παλινδρόμηση.

Εκτίμηση των συντελεστών της παλινδρόμησης

Η μέθοδος εκτίμησης της παλινδρόμησης με k ερμηνευτικές μεταβλητές (περιλαμβάνοντας τη σταθερά) βασίζεται, όπως και στην απλή παλινδρόμηση, στην ***ελαχιστοποίηση του Αθροίσματος των Τετραγώνων των Καταλοίπων*** (SSR), δηλαδή στην ελαχιστοποίηση της διακύμανσης των καταλοίπων.

Min SSR = Min V[e] = Min **Σ** e_i^2 **όπου** e_i **= εκτίμηση του** ε_i

Με τη λύση των k κανονικών εξισώσεων , έχουμε:

$$\hat{b} = (X'X)^{-1}X'Y$$

Όπου *(X'X)* : συμμετρική μήτρα (k , k) και $(X'X)^{-1}$ = αντίστροφη μήτρα

Με την Μ.Ε.Τ. , οι εκτιμητές των συντελεστών είναι αμερόληπτοι (unbiased)

$$V(\hat{b}_j) = \frac{\sigma_\varepsilon^2}{SST_j(1-R_j^2)}$$

όπου $SSTj = \sum_{i=1}^{n}(X_{i,j} - \overline{X}_j)^2$ και $R_j^2 =$ ο συντελεστής προσδιορισμού της παλινδρόμησης της X_j επί όλων των υπολοίπων ανεξαρτήτων μεταβλητών.

Συνολική Αξιολόγηση της παλινδρόμησης [01]

Συντελεστής Πολλαπλού Προσδιορισμού: R^2

Ο συντελεστής

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

επηρεάζεται από τον αριθμό παρατηρήσεων όπως και από τον αριθμό ερμηνευτικών μεταβλητών.

Προσαρμοσμένος Συντελεστής Προσδιορισμού: R^{*2}

$$R^{2*} = 1 - \frac{SSR/n-k}{SST/n-1} = 1 - \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_Y^2}$$

Όπου:

$\hat{\sigma}_\varepsilon^2$ = διακύμανση του διαταρακτικού όρου

$\hat{\sigma}_Y^2$ = διακύμανση της εξαρτημένης μεταβλητής

$$R^{2*} = 1 - \left[\frac{n-1}{n-k} \times (1 - R^2)\right]$$

SSE = Explained Sum of Squares (παλινδρόμηση)
SSR = Residual Sum of Squares (κατάλοιπα)

Συνολική Αξιολόγηση της παλινδρόμησης [02]

Έλεγχος του Fisher

Ο έλεγχος του Fisher εφαρμόζεται όπως και στην απλή παλινδρόμηση. Συμβάλλει στην αξιολόγηση της σημαντικότητας του υποδείγματος στο σύνολό του.

Τι μπορεί να υποδηλώνει ο έλεγχος του Fisher

Σε ορισμένες περιπτώσεις, ο έλεγχος του Fisher μας οδηγεί στην απόρριψη της υπόθεσης Ho: $b_1 = b_2 = \ldots = b_k = 0$ (δηλαδή υπάρχει τουλάχιστον ένας συντελεστής διαφορετικός από το 0).

Όμως ταυτόχρονα, ο σημειακός έλεγχος του student μπορεί να μας οδηγεί στο συμπέρασμα ότι, κανένας συντελεστής της παλινδρόμησης δεν είναι στατιστικά σημαντικός???

Γιατί αυτό το «παράλογο» - κατά πρώτη όψη - αποτέλεσμα;

Συνολική Αξιολόγηση της παλινδρόμησης [02]

Το παράλογο – κατά πρώτη όψη – αποτέλεσμα δεν είναι τόσο παράλογο!

Η σχετικά υψηλή τιμή του Fisher (Ισχύει H_1), ενώ παράλληλα όλα (ή τα περισσότερα) t-student είναι μη στατιστικά σημαντικά, μπορεί να συμβεί όταν υπάρχει συνδυασμένη επίδραση των ανεξαρτήτων μεταβλητών, δηλαδή όταν συσχετίζονται σε σημαντικό βαθμό μεταξύ τους.

Συνδυασμένη επίδραση ⟺ Συσχέτιση μεταξύ των X_j

↓

$X_j = F(X_1, \ldots, X_l, \ldots X_k)$ $l \neq k$

↓

$R_j^2 \neq 0$ με υψηλή τιμή

↓

$V(\hat{b}_j) = \frac{\sigma_\varepsilon^2}{SST_j(1-R_j^2)}$ Υψηλή όπως και το τυπικό σφάλμα σ_j του συντελεστή b_j

↓

$t = \left|\frac{\hat{b}_1}{\hat{\sigma}_{b_1}}\right|$ Η τιμή της t-στατιστικής είναι μικρή ➔ $\hat{b}_j = 0$

Μη Ξεχνάμε ότι, υψηλό τυπικό σφάλμα σημαίνει και μεγάλο Δ.Ε. για το συντελεστή

Σημειακός έλεγχος των k συντελεστών

Έλεγχος για κάθε συντελεστή ξεχωριστά

- ✓ Γνωστή πλέον διαδικασία: βασίζεται στην στατιστική του **t-student** .
- ✓ Στόχος του ελέγχου: επιβεβαίωση ότι **κάθε συντελεστής είναι διαφορετικός από το μηδέν:** κάθε ανεξάρτητη μεταβλητή επηρεάζει την εξαρτημένη μεταβλητή.
- ✓ Η **p-value** μας δίνει το βαθμό σημαντικότητας που πρέπει να επιλέξουμε για να δεχτούμε την υπόθεση Η1 (ο συντελεστής ≠ 0).

Ολοκληρώσαμε με τις βασικές – προ απαιτούμενες γνώσεις για την παλινδρόμηση.

Όλοι οι προαναφερόμενοι έλεγχοι έχουν σημασία μόνο και μόνο αν οι 4 Σημαντικές Υποθέσεις της γραμμικής παλινδρόμησης εξασφαλίζονται!

1Η Πρόβλημα: Πολυσυγγραμμικότητα

Συνδυασμένη επίδραση των ανεξαρτήτων μεταβλητών

Δύο βασικές αιτίες της πολυσυγγραμμικότητας

Structural multi collinearity

Προκύπτει από την ίδια την εξειδίκευση του υποδείγματος.
Αν πρέπει στο υπόδειγμα να ενσωματώσουμε μια ερμηνευτική X καθώς και την X^2, υπάρχει σαφής συσχέτιση.

Data multi collinearity

Προκύπτει από τα δεδομένα (τιμές των μεταβλητών) και δεν είναι τεχνούργημα (artefact) της ίδιας της εξειδίκευσης του υποδείγματος.

Ανεξαρτησία μεταξύ των ερμηνευτικών μεταβλητών

Πλήρη συσχέτιση

$$| r_{X1X2} | = 1$$

Πλήρης ή τέλεια πολυσυγγραμμικότητα.
Οι συντελεστές δεν μπορούν να εκτιμηθούν. Η μήτρα $[X'X]^{-1}$ δεν υπάρχει.

Καμία συσχέτιση

$$r_{X1X2} = 0 \qquad R^2_{Y,X1X2} = R^2_{Y,X1} + R^2_{Y,X2}$$

Κανένα πρόβλημα πολυσυγγραμμικότητας.
Οι συντελεστές μπορούν να εκτιμηθούν.

Μερική συσχέτιση

$$r_{X1X2} \neq 0 \quad και \quad | r_{X1X2} | \neq 1$$

Μερική η ατελής πολυσυγγραμμικότητα.
Οι συντελεστές μπορούν να εκτιμηθούν. Πρέπει όμως να ελέγξουμε σε ποιο βαθμό το πρόβλημα της πολυσυγγραμμικότητας είναι σοβαρό ή όχι. Σε ποιο βαθμό οι εκτιμήσεις των συντελεστών εκφράζουν καλά την καθαρή επιρροή των ερμηνευτικών μεταβλητών;

Ανεξαρτησία μεταξύ των ερμηνευτικών μεταβλητών

Πλήρης – τέλεια Πολυσυγγραμμικότητα

Όταν υπάρχει τέλεια γραμμική σχέση.

Υποθέστε ότι έχουμε το ακόλουθο μοντέλο:

$$Y=\beta_1+\beta_2 X_2+ \beta_3 X_3+e$$

Όπου οι τιμές του δείγματος για τα X_2 και X_3 είναι:

X_2	1	2	3	4	5	6
X_3	2	4	6	8	10	12

Παρατηρούμε ότι $X_3=2X_2$

Πηγή: Dimitrios Asteriou and Stephen G. Hall (2011), Applied Econometrics, Eds Palgrave Macmillan,

Ανεξαρτησία μεταξύ των ερμηνευτικών μεταβλητών

Ατελής Πολυσυγγραμμικότητα

- Η ατελής πολυσυγγραμμικότητα (ή σχεδόν πολυσυγγραμμικότητα) υπάρχει όταν οι ερμηνευτικές μεταβλητές σε μια εξίσωση συσχετίζονται, αλλά αυτή η συσχέτιση είναι λιγότερο από τέλεια.

- Αυτό εκφράζεται ως εξής:

$$X_3 = X_2 + v$$

Όπου *v* μια τυχαία μεταβλητή, η οποία μπορεί να παρατηρηθεί ως ένα «λάθος» στην ακριβή γραμμική σχέση.

Τι σημαίνει μερική – ατελής συσχέτιση;

1. Οι διακυμάνσεις και συνδιακυμάνσεις των συντελεστών είναι αρκετά μεγάλες.

Όσο πιο στενή είναι η συσχέτιση, τόσο έχουμε:

- ✓ **Υψηλό R^2 ➔ υψηλή τιμή της F-στατιστικής** (ενώ πιθανόν δεν θα έπρεπε)
- ✓ **Μεγάλα τυπικά σφάλματα.**
- ✓ **Επιρροή στο διάστημα εμπιστοσύνης των συντελεστών παλινδρόμησης.**
- ✓ **Κακή επιρροή στην t-στατιστική** που είναι μικρότερη από την τιμή που θα υπολογίζαμε εάν δεν υπήρχε συσχέτιση ανάμεσα στις ανεξάρτητες μεταβλητές.

  ➔ **ο στατιστικός έλεγχος των ατομικών συντελεστών δεν είναι ακριβής**

Τι σημαίνει μερική – ατελής συσχέτιση;

2. **Δεν μπορούμε να διαχωρίσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής ξεχωριστά.**

- ✓ Ορισμένοι συντελεστές είναι μη στατιστικά σημαντικοί: οι ανεξάρτητες μεταβλητές σε ατομική βάση δεν ερμηνεύουν σημαντικά την μεταβλητότητα της εξαρτημένης μεταβλητής

Η πολυσυγγραμμικότητα μπορεί να επηρεάσει τις τιμές των συντελεστών και τα πρόσημά τους.
Η πολυσυγγραμμικότητα μπορεί να συντελέσει έμμεσα σε λανθασμένη εξειδίκευση του υποδείγματος.

Τι σημαίνει μερική – ατελής συσχέτιση;

Τελικά η ***πολυσυγγραμμικότητα*** προκαλεί δύο βασικά προβλήματα:

- Οι εκτιμήσεις των συντελεστών γίνονται πολύ ευαίσθητοι σε οποιαδήποτε μικρή αλλαγή στο μοντέλο. Η αλλαγή έστω μιας ερμηνευτικής μεταβλητής μπορεί να αλλάξει σε σημαντικό βαθμό την τιμή των συντελεστών.

- Η πολυσυγγραμμικότητα μειώνει την ακρίβεια των συντελεστών εκτίμησης, πράγμα που αποδυναμώνει τη στατιστική ισχύ του υποδείγματος.

Όπως αναφέρει ο J. Frost:

> *You might not be able to trust the p-values to identify independent variables that are statistically significant*.

Πως μπορούμε να διαπιστώσουμε την ύπαρξη πολυσυγγραμμικότητας; [01]

1ος απλός τρόπος (ένδειξη)

- ✓ R^2 και F: υψηλές τιμές ενώ ταυτόχρονα,
- ✓ οι περισσότεροι ατομικοί συντελεστές δεν είναι στατιστικά σημαντικοί (μικρή τιμή της t-Στατιστικής).

➔ Η κατάσταση αυτή υποδηλώνει με βεβαιότητα την ύπαρξη πολυσυγγραμμικότητας στο υπόδειγμα.

Πως μπορούμε να διαπιστώσουμε την ύπαρξη πολυσυγγραμμικότητας; [02]

2ος απλός τρόπος (σοβαρή ένδειξη, όμως δεν φτάνει)

Σύγκριση μεταξύ:

- ✓ του **απλού συντελεστή συσχέτισης** (Pearson Correlation): **«ακαθάριστος» (Zero-order)** που δεν δίνει συστηματικά την πραγματική καθαρή επιρροή της μεταβλητής Xj στην εξαρτημένη Y.

και

- ✓ του **μερικού συντελεστής συσχέτισης (partial)** που δίνει την καθαρή επιρροή της μεταβλητής Xj όταν αφαιρείται η επιρροή των άλλων ερμηνευτικών μεταβλητών.

Πως μπορούμε να διαπιστώσουμε την ύπαρξη πολυσυγγραμμικότητας; [03]

3ος τρόπος (τελική επιβεβαίωση)

Tolerance factor = $TOL_j = 1 - R_j^2$

% της διακύμανσης της ερμηνευτικής X_j που ΔΕΝ ερμηνεύεται από τις άλλες ερμηνευτικές μεταβλητές.

όπου R_j^2 = συντελεστής πολλαπλού προσδιορισμού της παλινδρόμησης που αφορά τη μεταβλητή X_j σε σχέση με όλες τις υπόλοιπες ανεξάρτητες μεταβλητές, δηλαδή:

$$X_j = b_0 + b_1 X_1 + \ldots + b_{j-1} X_{j-1} + b_{j+1} X_{j+1} + \ldots + a_k X_k + \varepsilon_j$$

Αν η μεταβλητή Xj δεν συσχετίζεται με τις άλλες ανεξάρτητες μεταβλητές $R_j^2 = 0$, τότε :

$TOL_j = 1$

Αντίθετα, η συσχέτιση είναι τόσο έντονη, όσο ο TOL_j τείνει προς το 0.

Όταν $TOL_j < 50\%$:
πάνω από 50% της μεταβλητότητας της μεταβλητής Xj εξηγείται από τις άλλες ανεξάρτητες μεταβλητές του μοντέλου,
➔ ***ιδιαίτερα έντονο πρόβλημα***

Πως μπορούμε να διαπιστώσουμε την ύπαρξη πολυσυγγραμμικότητας; [04]

3ος τρόπος (συνέχεια)

Variance Inflation Factor = $$VIF_j = \frac{1}{1-R_j^2} = \frac{1}{TOL_j}$$ $$VIF_j \geq 1$$

Όταν η μεταβλητή X_j συσχετίζεται έντονα με τις άλλες μεταβλητές, τότε το R_j^2 τείνει προς το 1 και επομένως ο VIF_j τείνει προς το ∞.
Μεγάλες τιμές του VIF_j αναδεικνύουν έντονο πρόβλημα πολυσυγγραμμικότητας.

Στην περίπτωση όπου η μεταβλητότητα της X_j εξηγείται τουλάχιστον κατά 50% από τις άλλες ανεξάρτητες μεταβλητές (**$TOL_j > 0,5$**), τότε **$VIF_j > 2$**, με αποτέλεσμα η διακύμανση του συντελεστή να είναι μεγάλη.

Από τον ορισμό της, η ***διακύμανση του συντελεστή b_j*** είναι : $$Var(\hat{b}_j) = \frac{\hat{\sigma}^2}{[X^{*T}.X^*]} \times VIF_j$$

όπου $X^{*T}.X^*$ = Πίνακας συσχέτισης μεταξύ των ερμηνευτικών μεταβλητών

➔ ***Όσο μεγαλύτερος είναι ο δείκτης VIF_j, Τόσο μεγαλύτερη είναι η διακύμανση του συντελεστή, γεγονός που δεν είναι συμβατό με τις βασικές υποθέσεις της ΜΕΤ.***

Πως μπορούμε να διαπιστώσουμε την ύπαρξη πολυσυγγραμμικότητας; [04]

4ος τρόπος (Τελική επιβεβαίωση): *Condition Index*

Ο έλεγχος αυτός βασίζεται στις ιδιότητες του **Πίνακα συσχέτισης**: $C = X^{*T}.X^{*}$ μέγεθος (k, k).

Υπολογίζουμε τις **ιδιοτιμές** λ_j, οι οποίες προκύπτουν από την διαγωνιοποίηση του πίνακα συσχέτισης C: $|C-\lambda.I| = 0$.

Εφόσον έχουμε k ερμηνευτικές μεταβλητές, έχουμε ένα σύστημα με k εξισώσεις και k άγνωστες λ_j (j = 1, …k).

Από τις ιδιότητες του Πίνακα C: $\sum_{j=1}^{k} \lambda_j = k$

Όταν υπάρχει πολυσυγγραμμικότητα, ορισμένες ιδιοτιμές είναι πολύ μικρές και τείνουν προς το 0 . Κατά συνέπεια:

Αν $\lambda_j \rightarrow 0$, τότε $\lambda_{max} / \lambda_j \rightarrow \infty$. **Ο αριθμός των λ_j που τείνουν προς το 0 μας δίνει τον αριθμό των μεταβλητών που είναι προβληματικές.**

Ορίζεται ως **Condition Index**, τη στατιστική: $\Phi_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$

- $\Phi_j > 15$ πρόβλημα
- $\Phi_j > 30$ καταστροφικό πρόβλημα

SPSS: Στην εντολή Regression > Linear > Statistics Collinearity diagnostics

Το Output θα μας δώσει τον σχετικό πίνακα με τις τιμές των ιδιοτιμών (eigenvalues) και των δεικτών Φ_j

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η πολυσυγγραμμικότητα καθιστά δύσκολη την ερμηνεία των συντελεστών και μειώνει τη δύναμη του υποδείγματος ως προς τον εντοπισμό των ανεξάρτητων μεταβλητών που είναι πραγματικά στατιστικά σημαντικές. Πρόκειται για σοβαρό πρόβλημα, υπάρχει ωστόσο τρόπος επίλυσης του προβλήματος.

Το βασικό ζήτημα είναι ο εντοπισμός του βαθμού πολυσυγγραμμικότητας: όσο εντονότερος είναι ο βαθμός πολυσυγγραμμικότητας, τόσο σοβαρότερο είναι το πρόβλημα! Συνήθως θεωρείται ότι η περιορισμένη πολυσυγγραμμικότητα (TOL > 50% όμως TOL < 100% ή VIF < 2) δεν δημιουργεί σημαντικό πρόβλημα ως προς την ερμηνεία του υποδείγματος.

Όλα τα στατιστικά λογισμικά υπολογίζουν το VIF για κάθε ερμηνευτική μεταβλητή του υποδείγματος. Το VIF είναι ≥ 1 και δεν έχει ανώτατο όριο. VIF =1 σημαίνει καμία πολυσυγγραμμικότητα, VIF < 2: περιορισμένη πολυσυγγραμμικότητα η οποία δεν απαιτεί διόρθωση, VIF >2 και <5 συνιστάται διόρθωση και VIF ≥ 5: έντονο πρόβλημα που απαιτεί οπωσδήποτε διόρθωση.

Επίλυση Πολυσυγγραμμικότητας

Οι ευκολότερες διαδικασίες «θεραπείας» αυτών των προβλημάτων είναι:

(a) η παράλειψη μίας από τις συγγραμικές μεταβλητές

(b) η μετατροπή των υψηλά συσχετιζόμενων μεταβλητών σε ένα λόγο

(c) Ανάλυση κύριων συνιστωσών (principal component analysis)

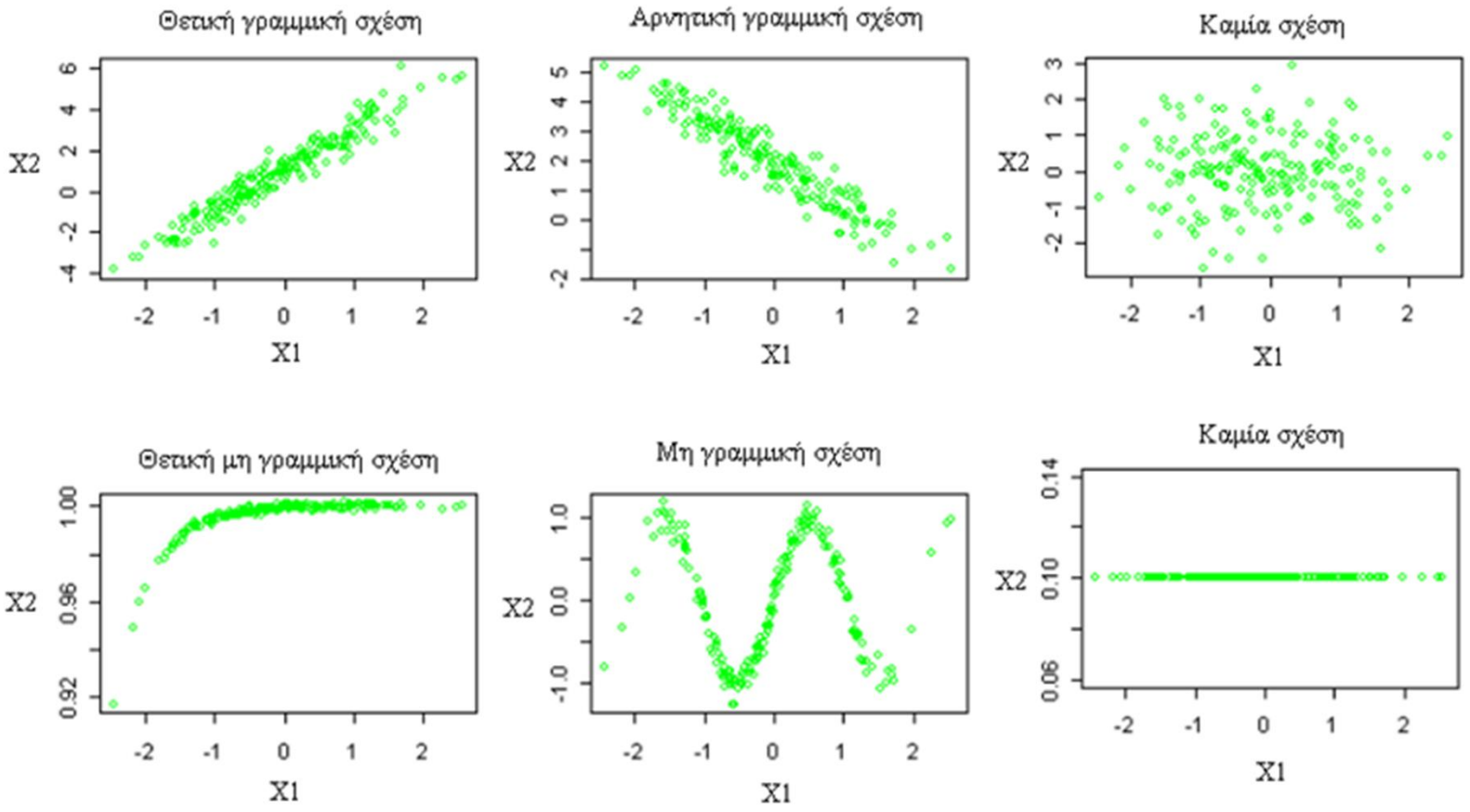
(d) η συλλογή περισσότερων δεδομένων (μακροπρόθεσμα δεδομένα)

(e) η μεγαλύτερη συχνότητα στα δεδομένα

Σε ορισμένες περιπτώσεις, η τυποποίηση των δεδομένων (standardization) μειώνει σημαντικά το βαθμό της πολυσυγγραμμικότητας. Η διαδικασία αυτή δεν αλλάζει την τιμή του R^2 ούτε του διορθωμένου R^2, διότι η πολυσυγγραμμικότητα δεν επηρεάζει την καλή «εξομάλυνση» των δεδομένων (goodness-of-fit).

ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ – ΑΝΕΞΑΡΤΗΣΙΑ ΜΕΤΑΞΥ ΤΩΝ ΕΡΜΗΝΕΥΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ

Ορισμένες μορφές σχέσεις μεταξύ 2 μεταβλητών



ΣΥΓΚΡΙΣΗ ΜΕΤΑΞΥ ΑΠΛΟΥ και ΜΕΡΙΚΟΥ ΣΥΝΤΕΛΕΣΤΗ ΣΥΣΧΕΤΙΣΗΣ

- **<u>Μερικός συντελεστής συσχέτισης</u>: (partial)** αναιρούμε την επίδραση που μπορεί να έχει μια τρίτη μεταβλητή (X_3) επάνω στις 2 πρώτες.

Απλός συντελεστής

Απαλλαγή της επίδρασης της X3 στις 2 άλλες μεταβλητές

$$r_{X1X2/X3} = \frac{r_{X1X2} - r_{X1X3} \times r_{X2X3}}{\sqrt{(1 - r_{X1X3}^2)} \times \sqrt{(1 - r_{X2X3}^2)}}$$

← Τυποποίηση έτσι ώστε $-1 \le r_{X1X2/X3} \le +1$

ΣΥΓΚΡΙΣΗ ΜΕΤΑΞΥ ΑΠΛΟΥ και ΜΕΡΙΚΟΥ ΣΥΝΤΕΛΕΣΤΗ ΣΥΣΧΕΤΙΣΗΣ

- **<u>Μερικοί συντελεστές συσχέτισης</u>: (partial) όταν έχουμε πάνω από 3 μεταβλητές.**

- Τύπος επανάληψης:

Μερικός συντελεστής X1X2, απαλλαγμένος από X3

Απαλλαγή της επίδρασης της X4

$$r_{X1X2/X3X4} = \frac{r_{X1X2/X3} - r_{X1X4/X3} \times r_{X2X4/X3}}{\sqrt{(1 - r^2_{X1X4/X3})} \times \sqrt{(1 - r^2_{X2X4/X3})}}$$

← Τυποποίηση έτσι ώστε $-1 \leq r_{X1X2/X3} \leq +1$

ΕΦΑΡΜΟΓΗ:

LECTURE_5_multiple linear regression.xls

Τα δεδομένα βρίσκονται στο φύλλο εργασίας: Data_Spss [A1:J136] ➔ 135 παρατηρήσεις

Θέμα: Συμπεριφορά των κατοίκων της Λάρισας σχετικά με την ανακύκλωση: **Id_beh**

Μεταβλητή	Τύπος μεταβλητής	Περιγραφή
Id_beh	Εξαρτημένη	Δείκτης συμπεριφοράς ως προς την ανακύκλωση (παίρνει τιμές μεταξύ 0 έως 100 όπου 0 = δεν συμμετέχει καθόλου ενώ 100 = σε καθημερινή βάση και για όλα τα είδη)
Sex	Ερμηνευτικές	Φύλο (διακριτή μεταβλητή όπου 0 = άνδρες και 1 = Γυναίκες)
AGE		*Ηλικία*
EDUC		*Επίπεδο εκπαίδευσης (1 έως 7)*
ID1		*Δείκτης αξιολόγησης της συμβολής της ανακύκλωσης στη βιώσιμη ανάπτυξη*
ID2		*Δείκτης αξιολόγησης της διάθεσης για ανακύκλωση*
ID3		*Δείκτης αξιολόγησης της προσβασιμότητας στις εγκαταστάσεις για ανακύκλωση*
ID4		*Δείκτης αξιολόγησης της συμβολής της ενημέρωσης*
ID5		*Δείκτης αξιολόγησης της επιρροής του άμεσου περιβάλλοντος*

Δείγμα:	150 ΚΑΤΟΙΚΟΙ ΛΑΡΙΣΑΣ
Περίοδος έρευνας	2014
ID1 έως ID5	Οι 5 δείκτες ID1 έως ID5 προέρχονται από την εφαρμογή της Ανάλυσης σε Κύριες Συνιστώσες, βασισμένη στις απαντήσεις πολλαπλών ερωτήσεων που τέθηκαν στους κατοίκους της Λάρισας.

Θέμα: Συμπεριφορά των κατοίκων της Λάρισας σχετικά με την ανακύκλωση: **Id_beh**

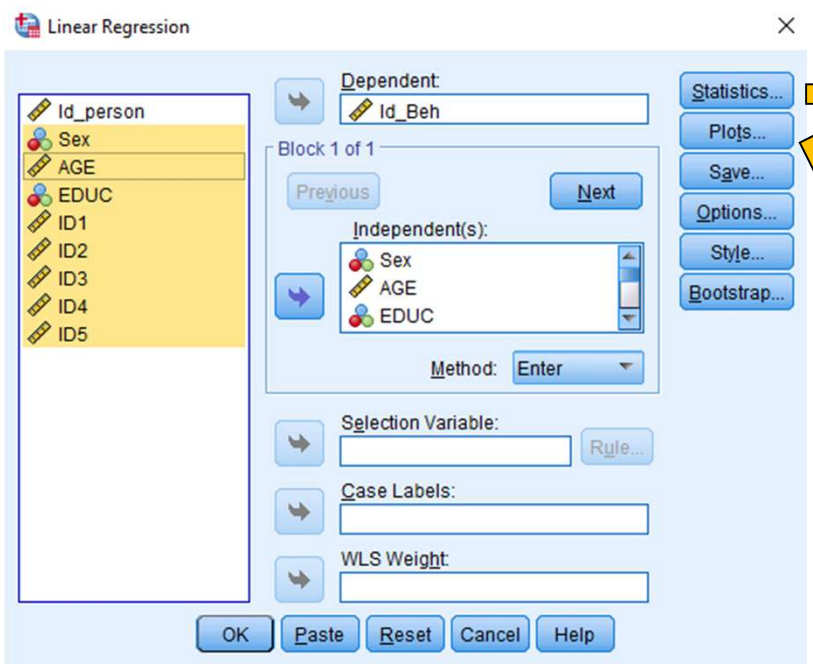
Οι ερμηνευτικές μεταβλητές αφορούν:

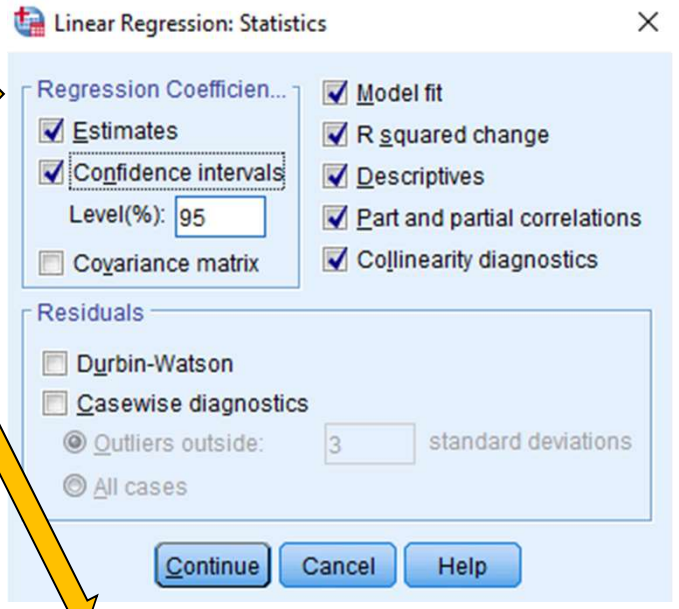
(α) ορισμένα κοινωνικά χαρακτηριστικά των κατοίκων (φύλο, ηλικία, επίπεδο εκπαίδευσης),
(β) πέντε «μετρήσεις» των «αντιλήψεων» των κατοίκων σε σχέση με την ανακύκλωση.

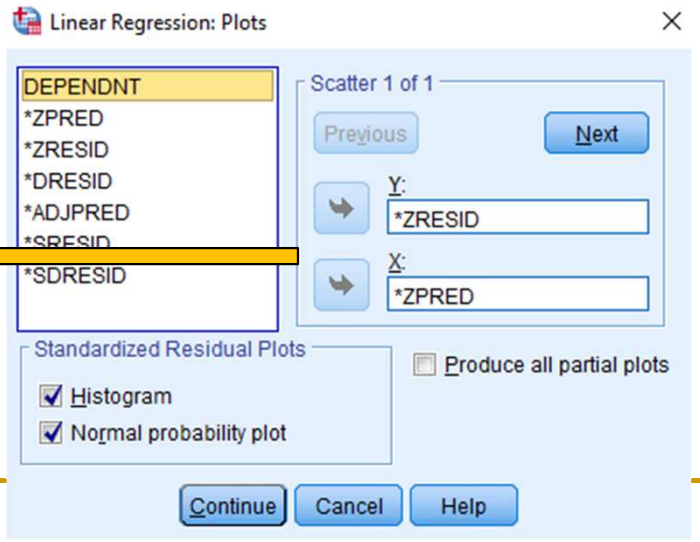
Συνολικά, έχουμε 8 ερμηνευτικές μεταβλητές.

Τα αποτελέσματα της πολλαπλής παλινδρόμησης δίνονται στις παρακάτω διαφάνειες.

Εφαρμογή με SPSS: Analyze, Regression, Linear







Διάγραμμα διασποράς των τιμών των τυποποιημένων κατάλοιπων (ZRESID) με βάση τις τυποποιημένες τιμές των εκτιμήσεων της εξαρτημένης μεταβλητής (ZPRED):

Έλεγχος της γραμμικότητας

Θέμα: Συμπεριφορά των κατοίκων της Λάρισας σχετικά με την ανακύκλωση: **Id_beh**

Correlations

		Id_Beh	Sex	AGE	EDUC	ID1	ID2	ID3	ID4	ID5
Pearson Correlation	Id_Beh	1,000	,023	-,571	,098	,380	-,373	,298	,290	,069
	Sex	,023	1,000	-,185	,106	-,100	,038	-,067	-,103	-,103
	AGE	-,571	-,185	1,000	-,345	-,298	,239	-,051	-,311	,197
	EDUC	,098	,106	-,345	1,000	-,043	,064	,033	,270	-,078
	ID1	,380	-,100	-,298	-,043	1,000	,001	,001	,046	-,022
	ID2	-,373	,038	,239	,064	,001	1,000	,064	-,091	,097
	ID3	,298	-,067	-,051	,033	,001	,064	1,000	,020	,031
	ID4	,290	-,103	-,311	,270	,046	-,091	,020	1,000	,044
	ID5	,069	-,103	,197	-,078	-,022	,097	,031	,044	1,000
Sig. (1-tailed)	Id_Beh	.	,396	,000	,130	,000	,000	,000	,000	,212
	Sex	,396	.	,016	,111	,124	,329	,218	,117	,117
	AGE	,000	,016	.	,000	,000	,003	,280	,000	,011
	EDUC	,130	,111	,000	.	,309	,231	,351	,001	,183
	ID1	,000	,124	,000	,309	.	,493	,494	,297	,400
	ID2	,000	,329	,003	,231	,493	.	,231	,146	,130
	ID3	,000	,218	,280	,351	,494	,231	.	,409	,362
	ID4	,000	,117	,000	,001	,297	,146	,409	.	,307
	ID5	,212	,117	,011	,183	,400	,130	,362	,307	.

?

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,758[a]	,575	,548	14,5714	,575	21,309	8	126	,000

a. Predictors: (Constant), ID5, ID1, ID3, ID4, ID2, Sex, EDUC, AGE

b. Dependent Variable: Id_Beh

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36194,798	8	4524,350	21,309	,000[b]
	Residual	26753,072	126	212,326		
	Total	62947,870	134			

a. Dependent Variable: Id_Beh

b. Predictors: (Constant), ID5, ID1, ID3, ID4, ID2, Sex, EDUC, AGE

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	70,511	7,979		8,837	,000					
	Sex	1,762	2,780	,039	,634	,527	,023	,056	,037	,888	1,126
	AGE	-,598	,107	-,412	-5,562	,000	-,571	-,444	-,323	,614	1,629
	EDUC	-,741	,979	-,049	-,757	,451	,098	-,067	-,044	,802	1,247
	ID1	5,609	1,379	,257	4,067	,000	,380	,341	,236	,848	1,180
	ID2	-6,854	1,424	-,297	-4,813	,000	-,373	-,394	-,280	,888	1,126
	ID3	6,444	1,292	,292	4,989	,000	,298	,406	,290	,981	1,019
	ID4	2,863	1,443	,126	1,984	,049	,290	,174	,115	,831	1,204
	ID5	3,766	1,321	,171	2,851	,005	,069	,246	,166	,940	1,064

a. Dependent _Beh

Ποια ερμηνεία μπορούν να δώσουμε στο πρόσημο της κάθε ερμηνευτικής μεταβλητής;
… όσο πιο ηλικιωμένος είναι ο κάτοικος τόσο λιγότερο ανακυκλώνει.

Ποιες ερμηνευτικές μεταβλητές συμβάλλουν στην ερμηνεία της διακύμανσης της εξαρτημένης μεταβλητής;

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

38,6% της διακύμανσης της μεταβλητής Age ερμηνεύεται από τις υπόλοιπες

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			linearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	rance	VIF
1	(Constant)	70,511	7,979		8,837	,000					
	Sex	1,762	2,780	,039	,634	,527	,023	,056	,037	,888	1,126
	AGE	-,598	,107	-,412	-5,562	,000	-,571	-,444	-,323	,614	1,629
	EDUC	-,741	,979	-,049	-,757	,451	,098	-,067	-,044	,802	1,247
	ID1	5,609	1,379	,257	4,067	,000	,380	,341	,236	,848	1,180
	ID2	-6,854	1,424	-,297	-4,813	,000	-,373	-,394	-,280	,888	1,126
	ID3	6,444	1,292	,292	4,989	,000	,298	,406	,290	,981	1,019
	ID4	2,863	1,443	,126	1,984	,049	,290	,174	,115	,831	1,204
	ID5	3,766	1,321	,171	2,851	,005	,069	,246	,166	,940	1,064

a. Dependent Variable: Id_Beh

Όλοι οι δείκτες TOL > 0,500 (> 50%) ➔ σχετική μικρή συσχέτιση της κάθε ερμηνευτικής μεταβλητής με όλες τις υπόλοιπες.

VIF < 2

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Εφαρμόζοντας την πολλαπλή παλινδρόμηση με :

- την μεταβλητή **AGE** ως εξαρτημένη και,
- οι μεταβλητές: SEX, EDUC, ID1, ID2, ID3, ID4 & ID5 ως ερμηνευτικές

Επιβεβαιώνουμε ότι:

ο συντελεστής προσδιορισμού = $R^2_{age} = 0{,}386 = 1 - TOL_{age}$

- Ο συντελεστής **VIF < 2** ➔ φαίνεται ότι η πολυσυγγραμμικότητα δεν είναι ιδιαίτερα προβληματική.

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions								
				(Constant)	Sex	AGE	EDUC	ID1	ID2	ID3	ID4	ID5
1	1	3,587	1,000	,00	,02	,01	,00	,00	,00	,00	,00	,00
	2	1,146	1,769	,00	,00	,00	,00	,04	,31	,10	,10	,22
	3	1,079	1,823	,00	,00	,00	,00	,12	,00	,12	,40	,15
	4	,998	1,896	,00	,00	,00	,00	,50	,08	,05	,08	,15
	5	,965	1,928	,00	,00	,00	,00	,12	,02	,64	,00	,19
	6	,809	2,105	,00	,00	,00	,00	,06	,50	,07	,22	,22
	7	,308	3,414	,00	,75	,05	,00	,00	,00	,00	,00	,04
	8	,092	6,235	,00	,16	,37	,30	,05	,01	,01	,19	,03
	9	,017	14,693	,99	,07	,57	,70	,11	,07	,00	,00	,00

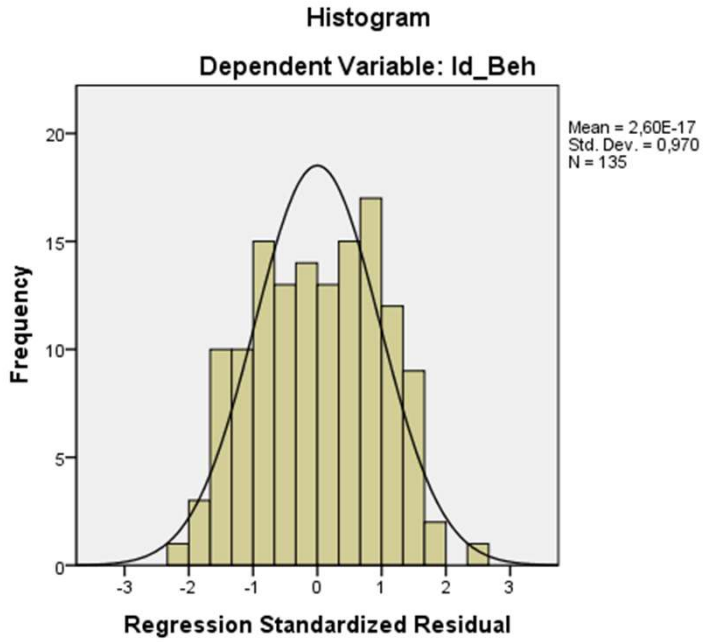
a. Dependent Variable: Id_Beh

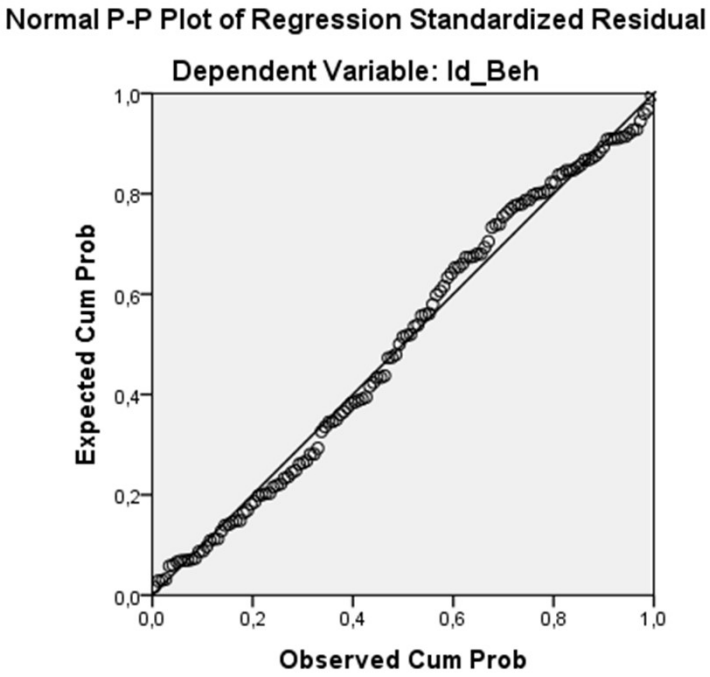
$$\sqrt{\frac{\lambda_{max}}{\lambda_2}} = \sqrt{\frac{3{,}587}{1{,}146}} = 1{,}769 < 15$$

Μόνο μια ιδιοτιμή τείνει πραγματικά προς το 0,
ενώ ο δείκτης Φ_j (Condition Index) είναι συστηματικά < 15

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

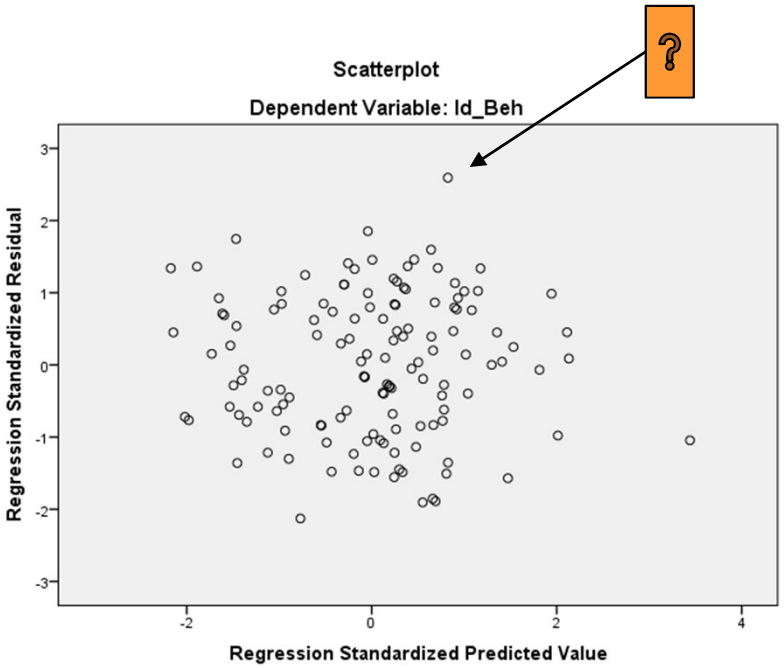
Έλεγχος της γραμμικότητας (Α)

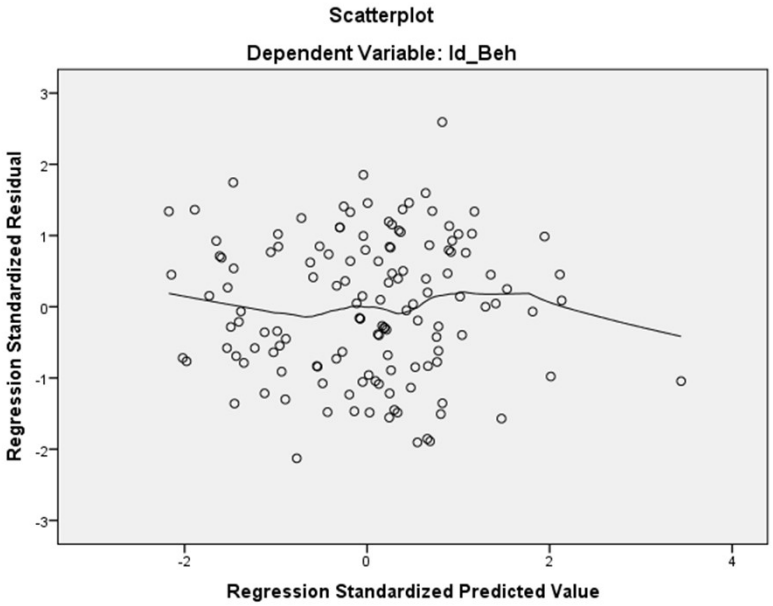




Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Έλεγχος της γραμμικότητας (B)





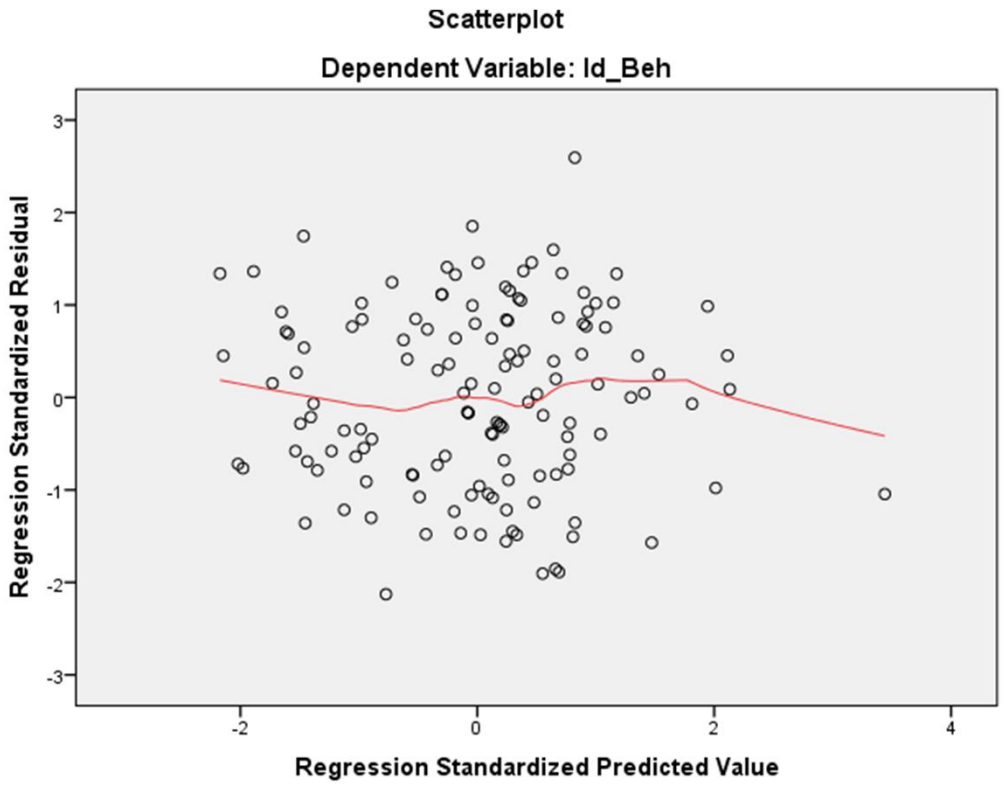
Τα τυποποιημένα κατάλοιπα κυμαίνονται μεταξύ -2 και +2 με μια εξαίρεση

Διπλό click μέσα στο διάγραμμα και επιλέξετε: ***Elements, Fit line at Total, Loess Curve, Apply***

Loess Curve (Local regression): μη παραμετρική μέθοδος εκτίμησης μιας καμπύλής μεταξύ των 2 μεταβλητών όπου κάθε τιμή της μεταβλητής X σταθμίζεται με τις πλησιέστερες τιμές της.

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Έλεγχος της γραμμικότητας (B)



Από την καμπύλη Loess, φαίνεται ότι η σχέση είναι ***κατά προσέγγιση γραμμική γύρω από το μηδέν.***

Φαίνεται ότι τα κατάλοιπα είναι τυχαία διάσπαρτα γύρω από το μηδέν

Αποτελέσματα για την συμπεριφορά σχετικά με την ανακύκλωση: **Id_beh**

Συμπεράσματα:

Η συμπεριφορά σχετικά με την ανακύκλωση (στην περίπτωση των κατοίκων της Λάρισας) δεν εξαρτάται ούτε από το φύλο ούτε από το επίπεδο εκπαίδευσης.

Αντιθέτως η ηλικία αποτελεί βασικό παράγοντα: όσο πιο νέος είναι ο κάτοικος τόσο πιο πολύ έχει – ceteris paribus - θετική συμπεριφορά.

Οι πέντε μεταβλητές που αναφέρονται στις αντιλήψεις των κατοίκων σχετικά με τη σημασία της ανακύκλωσης επηρεάζουν την συμπεριφορά.

Προέκυψε ότι, η «διάθεση» για ανακύκλωση επηρεάζει αρνητικά την συμπεριφορά !!!! Περίεργο αποτέλεσμα….

> Όμως, όπως φαίνεται στον πίνακα συσχέτισης, η μεταβλητή «διάθεση» συσχετίζεται θετικά και έντονα με την ηλικία. Μήπως αυτό σημαίνει ότι, οι ηλικιωμένοι έχουν σημαντική διάθεση όμως στην πραγματικότητα δεν ανακυκλώνουν.