



ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

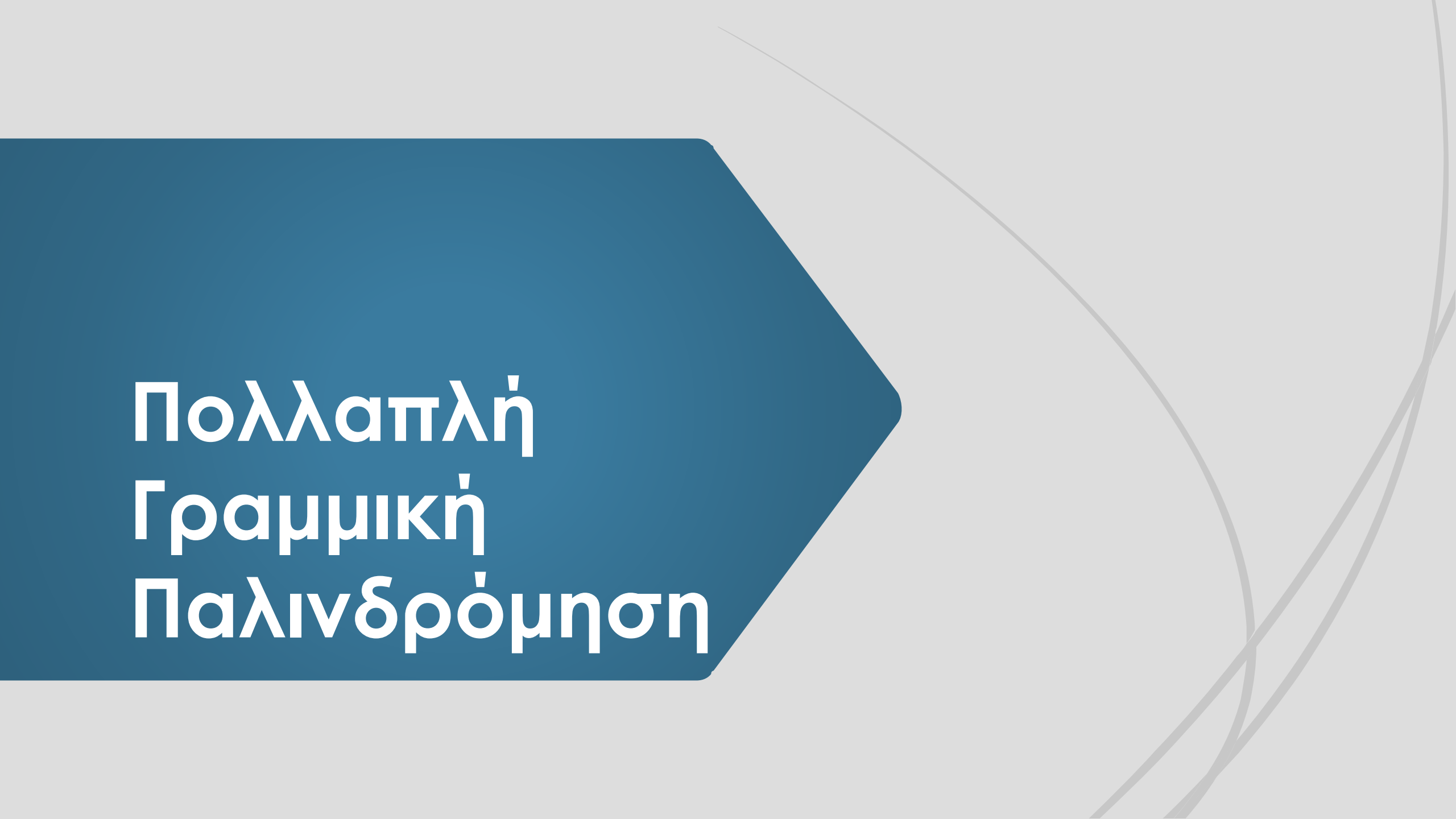
«Διοίκηση στη Ναυτική Επιστήμη και Τεχνολογία»



Image by David Mark from Pixabay

Διαχείριση μεγάλου όγκου δεδομένων στη Ναυτιλία και το Θαλάσσιο Περιβάλλον

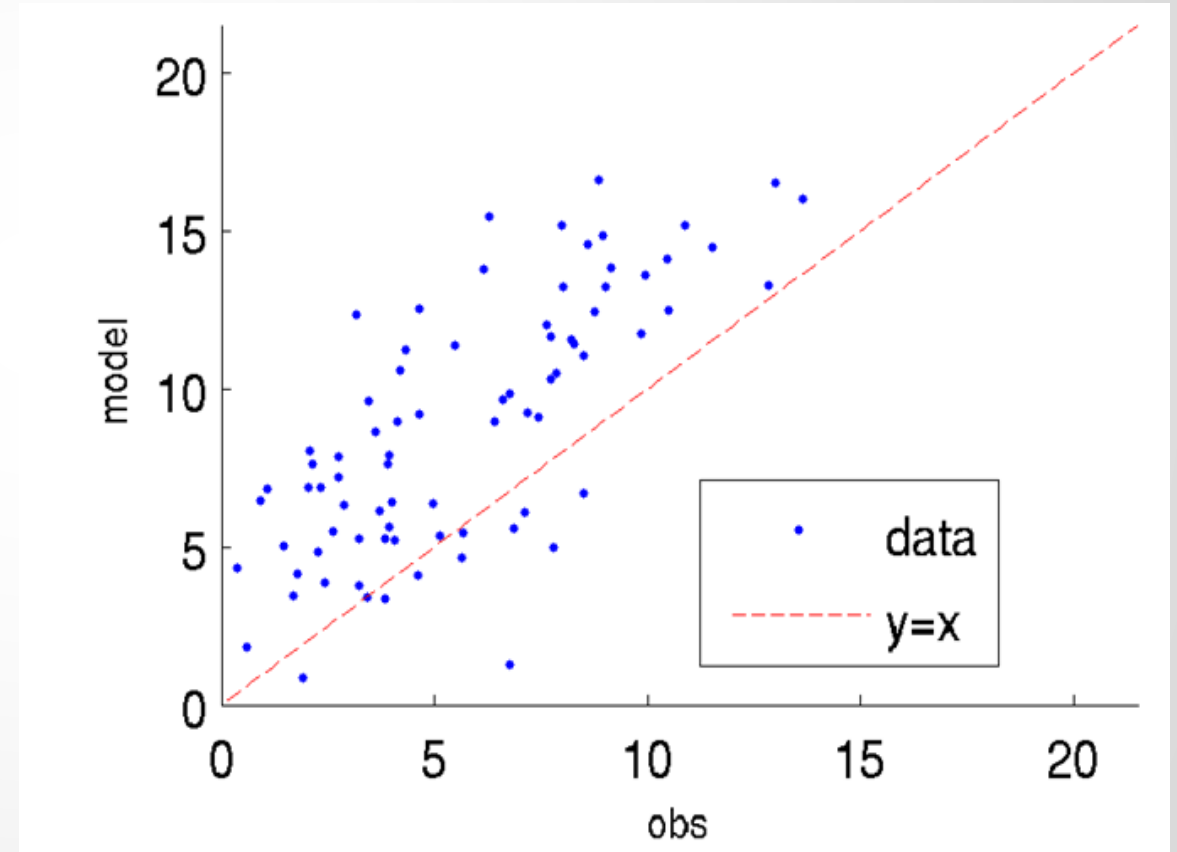
Διδάσκοντες: Γ. Γαλάνης, Δ. Κουλουμπού



Πολλαπλή Γραμμική Παλινδρόμηση

Από την Απλή στη Πολλαπλή Γραμμική Παλινδρόμηση

- Το υπόδειγμα της απλής γραμμικής παλινδρόμησης, αναφέρεται σε σχέσεις που περιλαμβάνουν μία μόνο ερμηνευτική μεταβλητή.



Από την Απλή στη Πολλαπλή Γραμμική Παλινδρόμηση

- Επέκταση της απλής παλινδρόμησης για δύο ή περισσότερες ανεξάρτητες μεταβλητές, αποτελεί η **πολλαπλή γραμμική παλινδρόμηση (Multiple Linear Regression)**.

Από την Απλή στη Πολλαπλή Γραμμική Παλινδρόμηση

- Από δεδομένα βιβλιογραφίας διαφαίνεται ευρεία χρήση του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης σε λύσεις επιχειρηματικών προβλημάτων.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης χρησιμοποιεί $k \geq 2$ ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k$ για την πρόβλεψη της εξαρτημένης μεταβλητής Y για την οποία υπάρχει ενδιαφέρον.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- ➡ Δηλαδή, η πολλαπλή ανάλυση παλινδρόμησης μας δίνει πληροφορίες για το πως μια μεταβλητή σχετίζεται μ' άλλες (ανεξάρτητες) μεταβλητές και μας δίνει την εξίσωση γραμμικής παλινδρόμησης που επιτρέπει τον υπολογισμό (πρόβλεψη) της άγνωστης μεταβλητής αν οι τιμές των υπολοίπων είναι γνωστές.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Ας υποθέσουμε ότι έχουμε στη διάθεση μας n παρατηρήσεις από τις τυχαίες μεταβλητές $Y, X_1, X_2, \dots, X_k$ όπου όλες θεωρούνται συνεχείς μεταβλητές.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Σκοπός της πολλαπλής γραμμικής παλινδρόμησης είναι η δημιουργία ενός γραμμικού μοντέλου της μορφής

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + \varepsilon$$

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

Όπου:

- Οι $X_1, X_2, \dots, X_k$ ονομάζονται **ανεξάρτητες** ή **ερμηνευτικές μεταβλητές** (independent or explanatory variables).

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Η Y ονομάζεται **εξαρτημένη μεταβλητή** (dependent variable) και είναι αυτή η μεταβλητή της οποίας τις τιμές προσπαθούμε να περιγράψουμε (προβλέψουμε) ως συνάρτηση (με γραμμικό τρόπο) των ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_k$.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Στο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + \varepsilon$$

- b_0 είναι ο **σταθερός όρος** της Y στην επιφάνεια της παλινδρόμησης.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Τα $b_i, i = 1, 2, \dots, k$ είναι οι συντελεστές της παλινδρόμησης.
- ε είναι ο όρος σφάλματος του μοντέλου.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Συνεπώς, το μοντέλο θα αποτελείται από τα ακόλουθα στοιχεία:
 - Τις παρατηρούμενες τιμές $Y_1, Y_2, \dots, Y_n$ που έχουν καταγραφεί στο διαθέσιμο σύνολο των δεδομένων.
 - Τις αντίστοιχες παρατηρούμενες τιμές $X_{1,i}, X_{2,i}, \dots, X_{k,i}$ για κάθε Y_i , $i = 1, 2, \dots, n$.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Τις προβλεπόμενες τιμές $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$ της Y όπως εκτιμώνται από το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης που έχει προσαρμοσθεί.

Δηλαδή

$$\hat{Y}_i = b_0 + b_1X_{1,i} + b_2X_{2,i} + \dots + b_kX_{k,i}$$

για κάθε $i = 1, 2, \dots, n$.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

- Τα σφάλματα ή κατάλοιπα (residuals) του μοντέλου ε_i , όπου

$$\varepsilon_i = Y_i - \hat{Y}_i \text{ για κάθε } i = 1, 2, \dots, n$$

- Τις παραμέτρους $b = (b_0, b_1, b_2, \dots, b_k)^T$ του μοντέλου.

Το μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης

➤ Τη διακύμανση $\hat{\sigma}^2$ του μοντέλου, όπου

$$\hat{\sigma}^2 = \frac{1}{n - k - 1} \sum_{i=1}^n \varepsilon_i^2 = \frac{1}{n - k - 1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Πολλαπλή Γραμμική Παλινδρόμηση – Βασικές Υποθέσεις

- Όπως και στο απλό γραμμικό υπόδειγμα έτσι και στο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης υπάρχουν κάποιες υποθέσεις οι οποίες πρέπει να ισχύουν για όλες τις παρατηρήσεις.

Αυτές είναι οι εξής:



Πολλαπλή Γραμμική Παλινδρόμηση – Βασικές Υποθέσεις

- Για κάθε παρατήρηση ο όρος του σφάλματος ε_i είναι κανονικά κατανοημένος με μέση τιμή 0 και τυπική απόκλιση σ και είναι ανεξάρτητος από τους όρους του σφάλματος που σχετίζονται με όλες τις άλλες παρατηρήσεις.

Δηλαδή,

$$\varepsilon_i \sim N(0, \sigma^2) \text{ για κάθε } i = 1, 2, \dots, n$$

ανεξάρτητα από τα άλλα σφάλματα.

Πολλαπλή Γραμμική Παλινδρόμηση – Βασικές Υποθέσεις

- Οι ερμηνευτικές μεταβλητές δεν είναι στοχαστικές. Οι τιμές τους παραμένουν σταθερές και δεν είναι όλες ίσες μεταξύ τους.
- Δεν υπάρχουν ακριβείς γραμμικές σχέσεις ανάμεσα στις ερμηνευτικές μεταβλητές.
- Ο αριθμός των παρατηρήσεων του δείγματος είναι μεγαλύτερος από τον αριθμό των συντελεστών του υποδείγματος που θέλουμε να εκτιμήσουμε.

Κριτήριο Ελάχιστων Τετραγώνων

- Οι συντελεστές του γραμμικού μοντέλου υπολογίζονται με το ίδιο κριτήριο όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης.
- Δηλαδή με το κριτήριο των ελάχιστων τετραγώνων ή με άλλα λόγια ελαχιστοποιώντας τη διακύμανση των σφαλμάτων του μοντέλου.

Κριτήριο Ελάχιστων Τετραγώνων

$$\begin{aligned} & \min_{b_0, b_1, b_2, b_\kappa} \sum_{i=1}^k (Y_i - \hat{Y}_i)^2 = \\ & = \min_{b_0, b_1, b_2, b_\kappa} \sum_{i=1}^k \left(Y_i - (b_0 + b_1 X_{1,i} + b_2 X_{2,i} + \cdots + b_\kappa X_{\kappa,i}) \right)^2 \end{aligned}$$

Χρήση Υπολογιστικών Προγραμμάτων

- Βέβαια όσο περισσότερες είναι οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k$, τόσο πιο σύνθετο είναι το πρόβλημα υπολογισμού των $b_0, b_1, b_2, \dots, b_k$.
- Η χρήση υπολογιστικών προγραμμάτων όπως Excel, Octave ή Matlab, SPSS, Jasp και άλλων είναι σχεδόν απαραίτητη σε τέτοιους σύνθετους υπολογισμούς.

Ερμηνεία Συντελεστών

- Η σταθερά b_0 ερμηνεύεται ως η αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y αν όλες οι ανεξάρτητες μεταβλητές είναι 0.
- Ο συντελεστής παλινδρόμησης b_j εκφράζει τη μεταβολή της εξαρτημένης μεταβλητής Y για κάθε μοναδιαία μεταβολή της ανεξάρτητης μεταβλητής X_j αν όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές.

Αξιολόγηση του Μοντέλου

- Η αξιολόγηση του μοντέλου μπορεί να πραγματοποιηθεί με έναν από τους παρακάτω τρόπους
 - **Τυπικό σφάλμα εκτίμησης, S_E**
 - **Συντελεστής προσδιορισμού R^2 και προσαρμοσμένος συντελεστής προσδιορισμού $\bar{R}^2$**

Τυπικό Σφάλμα Εκτίμησης

- Για ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης το **τυπικό σφάλμα εκτίμησης** δίνεται από τον τύπο

$$S_E = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - k - 1}}$$

Όπου n είναι το μέγεθος του δείγματος και k ο αριθμός των ανεξάρτητων μεταβλητών στο μοντέλο.

Τυπικό Σφάλμα Εκτίμησης

- Η μικρότερη τιμή που μπορεί να πάρει το S_E είναι το 0, δηλαδή όταν όλα τα σημεία του δείγματος ανήκουν στην ευθεία παλινδρόμησης.
- Γενικά όταν το S_E είναι μικρό η προσαρμογή θεωρείται καλή και το γραμμικό μοντέλο ενδέχεται να ένα αποτελεσματικό προγνωστικό και ερμηνευτικό εργαλείο.

Τυπικό Σφάλμα Εκτίμησης

- Όταν το S_E είναι μεγάλο η προσαρμογή δεν θεωρείται καλή και το γραμμικό μοντέλο είναι φτωχό στην προγνωστική του ικανότητα. Ο ερευνητής σε αυτήν την περίπτωση θα πρέπει να το βελτιώσει ή να το απορρίψει.
- Γενικά το τυπικό σφάλμα εκτίμησης S_E δεν μπορεί να χρησιμοποιηθεί ως απόλυτο μέτρο της χρησιμότητας του μοντέλου.

Τυπικό Σφάλμα Εκτίμησης

- Ωστόσο το τυπικό σφάλμα εκτίμησης S_E είναι χρήσιμο στη σύγκριση μοντέλων. Εάν ο ερευνητής έχει τη δυνατότητα να επιλέξει από πολλά μοντέλα, θα πρέπει γενικά να επιλέξει αυτό με την μικρότερη τιμή του S_E .

Συντελεστής Προσδιορισμού R^2

- Στην πολλαπλή γραμμική παλινδρόμηση, ο συντελεστής πολλαπλού προσδιορισμού, ορίζεται ως το ποσοστό της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής Y που εξηγείται από την εξίσωση της πολλαπλής παλινδρόμησης, το οποίο οφείλεται στις επιδράσεις όλων μαζί των ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_k$.

Συντελεστής Προσδιορισμού R^2

Ο συντελεστής προσδιορισμού του μοντέλου δίνεται από τη σχέση

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

ή ισοδύναμα,

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Συντελεστής Προσδιορισμού R^2

Όπου

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

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

Συντελεστής Προσδιορισμού R^2

- Υπενθυμίζεται ότι ισχύει

$$SST = SSR + SSE.$$

- Ο συντελεστής προσδιορισμού R^2 εκφράζει το ποσοστό της μεταβλητότητας της μεταβλητής Y , που ερμηνεύονται από τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k$.
- Ισχύει $0 \leq R^2 \leq 1$

Προσαρμοσμένος Συντελεστής Προσδιορισμού $\bar{R}^2$

- Γενικά ο συντελεστής προσδιορισμού R^2 πρέπει να χρησιμοποιείται όταν το δείγμα μας είναι αρκετά μεγάλο σε σχέση με τον αριθμό των παραμέτρων μας.

Προσαρμοσμένος Συντελεστής Προσδιορισμού $\bar{R}^2$

- Αν όμως δεν συμβαίνει αυτό, τότε χρησιμοποιούμε τον προσαρμοσμένο (adjusted) συντελεστή πολλαπλού προσδιορισμού ο οποίος συμβολίζεται με $\bar{R}^2$ και έχει τον εξής τύπο :

Προσαρμοσμένος Συντελεστής Προσδιορισμού $\bar{R}^2$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - (k + 1)}$$

Όπου:

n : Το πλήθος των παρατηρήσεων

k : Το πλήθος των ανεξάρτητων (ερμηνευτικών)
μεταβλητών

Προσαρμοσμένος Συντελεστής

Προσδιορισμού $\bar{R}^2$

- Μια από τις βασικές διαφορές των R^2 , $\bar{R}^2$ είναι ότι με την αύξηση του αριθμού των ανεξάρτητων μεταβλητών, η τιμή του συντελεστή προσδιορισμού R^2 πάντοτε αυξάνει, ενώ η τιμή του διορθωμένου συντελεστή προσδιορισμού $\bar{R}^2$ ενδέχεται να αυξηθεί ή να μειωθεί.

Προσαρμοσμένος Συντελεστής Προσδιορισμού $\bar{R}^2$

- Ο προσαρμοσμένος συντελεστής προσδιορισμού είναι ένα χρήσιμο μέτρο, διότι όταν ο αριθμός των ανεξάρτητων μεταβλητών k είναι μεγάλος σε σχέση με το μέγεθος του δείγματος n , ο συντελεστής προσδιορισμού R^2 μπορεί να είναι μη ρεαλιστικά υψηλή. Για να αποφευχθεί η δημιουργία λανθασμένης εντύπωσης, συχνά υπολογίζουμε τον προσαρμοσμένο συντελεστής προσδιορισμού $\bar{R}^2$.

Αξιολόγηση Μοντέλου

Σε γενικές γραμμές μία ιδέα για την αξιολόγηση του μοντέλου μας δίνει και ο παρακάτω πίνακας.

SSE	S_E	$\bar{R}^2$	R^2	Αξιολόγηση Μοντέλου
0	0	1	1	Τέλειο
Μικρό	Μικρό	Κοντά στο 1	Κοντά στο 1	Καλό
Μεγάλο	Μεγάλο	Κοντά στο 0	Κοντά στο 0	Κακό

Περίπτωση δύο Ανεξάρτητων Μεταβλητών

- Η πιο απλή περίπτωση της πολλαπλής γραμμικής παλινδρόμησης είναι αυτή των δύο ανεξάρτητων μεταβλητών X_1, X_2 με εξίσωση

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

Δηλαδή,

$$\hat{Y}_i = b_0 + b_1X_{1,i} + b_2X_{2,i}$$

Για κάθε $i = 1, 2, \dots, n$.

Περίπτωση δύο Ανεξάρτητων Μεταβλητών

- Στην περίπτωση αυτή, αποδεικνύεται ότι οι τιμές των συντελεστών b_0, b_1, b_2 προκύπτουν από την επίλυση του συστήματος:

Περίπτωση δύο Ανεξάρτητων Μεταβλητών – Κανονικές Εξισώσεις

$$\left\{ \begin{array}{l} nb_0 + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n (x_{1i})^2 + b_2 \sum_{i=1}^n x_{1i}x_{2i} = \sum_{i=1}^n x_{1i}y_i \\ b_0 \sum_{i=1}^n x_{2i} + b_1 \sum_{i=1}^n x_{1i}x_{2i} + b_2 \sum_{i=1}^n (x_{2i})^2 = \sum_{i=1}^n x_{2i}y_i \end{array} \right.$$

Παράδειγμα 1

- Ένα μεσιτικό γραφείο θέλοντας να επενδύσει σε αγορά διαμερισμάτων σε συγκεκριμένη περιοχή της Ελλάδας έκανε μια έρευνα αγοράς σε 11 διαμερίσματα και κατέγραψε για κάθε ένα από αυτά την τιμή πώλησης σε χιλιάδες €, το εμβαδόν σε τετραγωνικά μέτρα και την παλαιότητα σε έτη.
- Στον επόμενο πίνακα παρουσιάζονται τα στοιχεία της έρευνας.

Παράδειγμα 1

Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	Εμβαδόν σε τ.μ	Παλαιότητα σε έτη
422	200	4
387	171	12
379	145	8
430	176	1
396	193	7
352	120	32
434	190	1
397	139	1
373	154	13
419	189	3
384	159	7

Παράδειγμα 1

- Παρατηρώντας τις τιμές του πίνακα διαπιστώνουμε, ότι η τιμή του διαμερίσματος εξαρτάται τόσο από το εμβαδόν του όσο και από την παλαιότητα του.
- Προκειμένου να έχουμε μία πιο καθαρή εικόνα, αλλά και για να προβλέψουμε την τιμή ενός διαμερίσματος όταν είναι γνωστό το εμβαδό και η παλαιότητα του θα δημιουργήσουμε ένα (γραμμικό) μοντέλο που θα περιγράφει τη σχέση της τιμής του διαμερίσματος με το εμβαδόν και την παλαιότητα του.

Παράδειγμα 1

- Σημαντικό επίσης είναι να βρούμε και την ερμηνευτική ικανότητα των ανεξάρτητων μεταβλητών.
- Για τη δημιουργία αυτού του μοντέλου αρχικά ονομάζουμε τις μεταβλητές.

Έτσι έχουμε:

Παράδειγμα 1

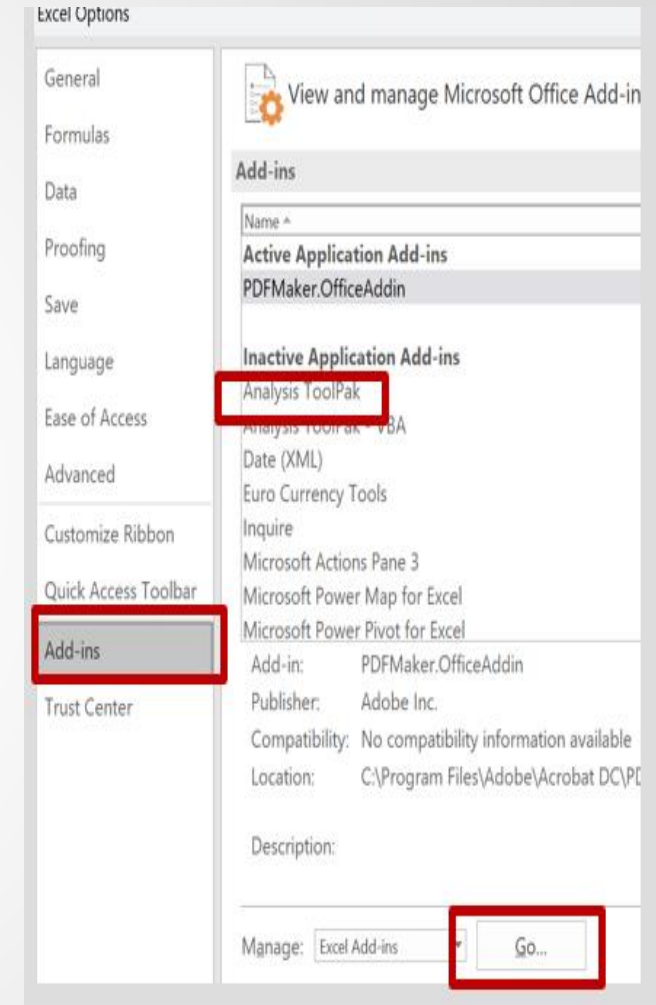
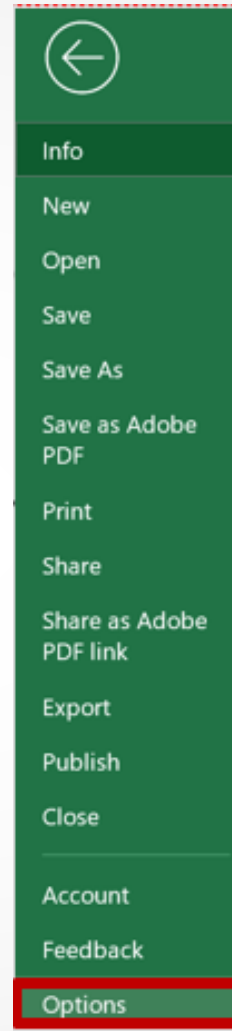
- Y : Τιμή διαμερίσματος σε χιλιάδες €.
- X_1 : Εμβαδόν διαμερίσματος σε τ. μ.
- X_2 : Παλαιότητα διαμερίσματος σε έτη.

Παράδειγμα 1 – Υλοποίηση Μέσω **Excel**

- Για την ανάλυση του μοντέλου πολλαπλής γραμμικής παλινδρόμησης μέσω **excel**, χρειάζεται να εγκαταστήσουμε στο excel το πακέτο **Analysis Toolpak**.
- Η διαδικασία αυτή γίνεται ως εξής:
- Από το excel ακολουθούμε

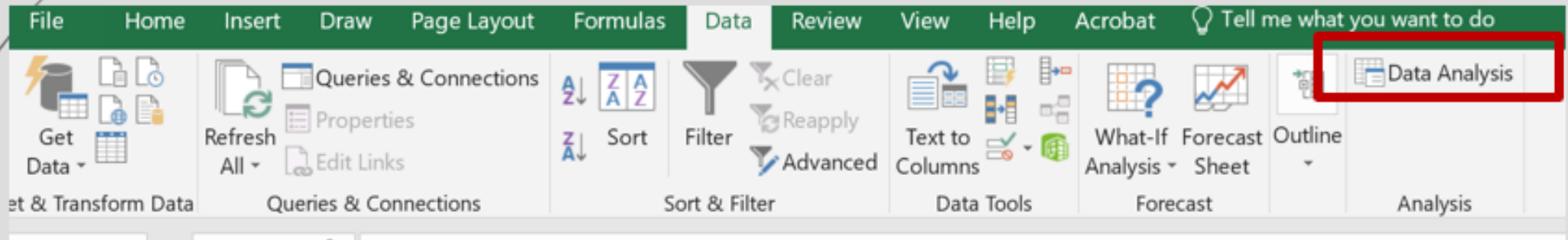
File → Options → Add – ins → Analysis Toolpak → Go

Παράδειγμα 1 – Υλοποίηση Μέσω MscExcel



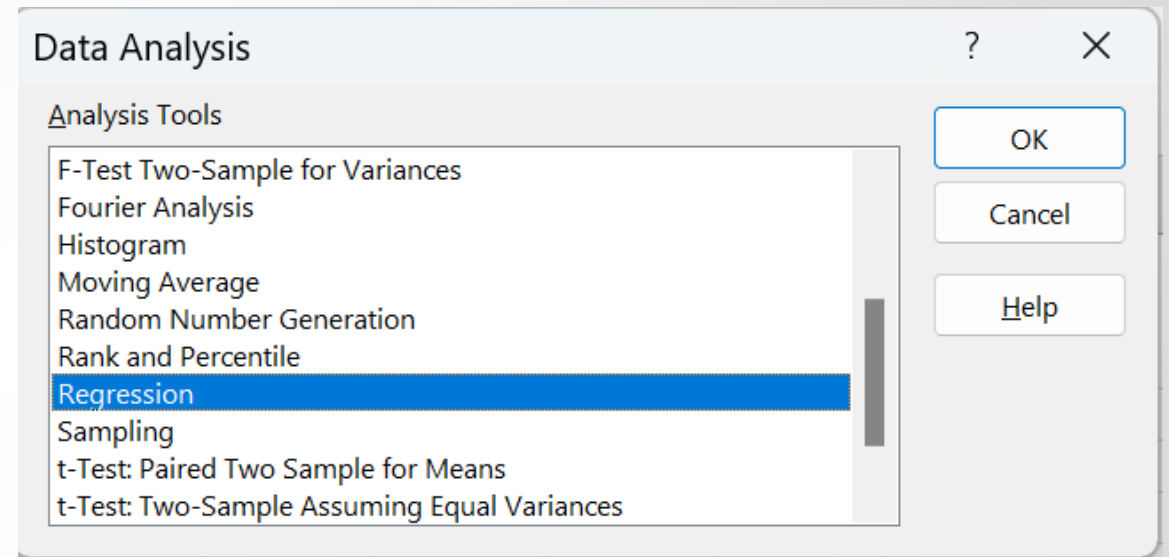
Παράδειγμα 1 – Υλοποίηση Μέσω Excel

- Η επιλογή Data Analysis είναι πλέον διαθέσιμη στο **excel**.



Παράδειγμα 1 – Υλοποίηση Μέσω Excel

- Μέσω της επιλογής **Data Analysis** του **Excel**, επιλέγουμε **Regression** και **ok**



Παράδειγμα 1 – Υλοποίηση Μέσω **Excel**

- Παρακάτω βλέπουμε τον συγκεντρωτικό πίνακα που δίνει το **excel** για το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης.

Παράδειγμα 1 – Υλοποίηση Μέσω Excel

SUMMARY OUTPUT									
<i>Regression Statistics</i>									
Multiple R	0,89752								
R Square	0,80555								
Adjusted R Square	0,75694								
Standard Error	12,8172								
Observations	11								
<i>ANOVA</i>									
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>				
Regression	2	5444,49	2722,25	16,5708	0,00143				
Residual	8	1314,24	164,28						
Total	10	6758,73							
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95,0%</i>	<i>Upper 95,0%</i>	
Intercept	332,621	37,5284	8,86317	2,1E-05	246,08	419,162	246,08	419,162	
Εμβαδόν σε τ.μ	0,46389	0,20457	2,26762	0,05309	-0,00785	0,93564	-0,00785	0,93564	
Παλαιότητα σε έτη	-1,54538	0,5844	-2,64439	0,02951	-2,89301	-0,19775	-2,89301	-0,19775	

Παράδειγμα 1 – Υλοποίηση Μέσω Excel

Από το πρώτο πλαίσιο του πίνακα βλέπουμε ότι ο συντελεστής προσδιορισμού (**R Square**) είναι

$$R^2 = 0,8055 = 80,55\%$$

Ενώ ο προσαρμοσμένος (adjusted) συντελεστής πολλαπλού προσδιορισμού (**Adjusted R Square**)

$$\bar{R}^2 = 0,7569 = 75,69\%$$

Regression Statistics	
Multiple R	0,89752
R Square	0,80555
Adjusted R Square	0,75694
Standard Error	12,8172
Observations	11

Παράδειγμα 1 – Υλοποίηση Μέσω Excel

- Δηλαδή το 75,69% της μεταβλητότητας της τιμής πώλησης των διαμερισμάτων, ερμηνεύεται από κοινού από το εμβαδόν και την παλαιότητα του.

Παράδειγμα 1 – Υλοποίηση Μέσω Excel

- Το τυπικό σφάλμα εκτίμησης είναι (Standard Error) είναι $S_E = 12,817$.

<i>Regression Statistics</i>	
Multiple R	0,89752
R Square	0,80555
Adjusted R Square	0,75694
Standard Error	12,8172
Observations	11

Παράδειγμα 1 – Υλοποίηση Μέσω Excel

- Το τελευταίο πλαίσιο του πίνακα μας δίνει την εξίσωση της παλινδρόμησης που είναι

$$Y = 332,62 + 0,4638X_1 - 1,5453X_2 + \varepsilon$$

Δηλαδή,

$$\hat{Y}_i = 332,62 + 0,4638X_{1,i} - 1,5453X_{2,i}$$

	Coefficients	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95,0%</i>	<i>Upper 95,0%</i>
Intercept	332,6208172	37,52844531	8,86317	2,1E-05	246,08	419,162	246,08	419,162
Εμβαδόν σε τ.μ	0,46389432	0,204572939	2,26762	0,05309	-0,00785	0,93564	-0,00785	0,93564
Παλαιότητα σε έτη	-1,54538158	0,58440076	-2,64439	0,02951	-2,89301	-0,19775	-2,89301	-0,19775

Παράδειγμα 1

- Η σταθερά $b_0 = 332,62$ μας λέει ότι δεδομένου ότι ένα σπίτι είναι 0τμ, και η παλαιότητα του είναι 0 έτη τότε η αναμενόμενη τιμή του είναι **332,62** χιλιάδες €. (Στο συγκεκριμένο παράδειγμα δεν έχει νόημα να ερμηνεύσουμε τη σταθερά, διότι ένα σπίτι με μηδενικό εμβαδό δεν υφίσταται).

Παράδειγμα 1

- Ο συντελεστής μερικής κλίσης $b_1 = 0,4638$ μας λέει ότι για κάθε ένα τ.μ. αύξησης του εμβαδού του διαμερίσματος και δεδομένου ότι η παλαιότητα του παραμένει σταθερή, η τιμή του αναμένεται να αυξάνεται κατά 0,4638 χιλιάδες € (463,8€).

Παράδειγμα 1

- Ο συντελεστής μερικής κλίσης $b_2 = -1,5453$ μας λέει ότι για κάθε ένα έτος αύξησης της παλαιότητας του διαμερίσματος και δεδομένου ότι το εμβαδόν του παραμένει σταθερό, η τιμή του αναμένεται να μειώνεται κατά 1,5453 χιλιάδες € (1545,3€).

Έλεγχος για τη Σημαντικότητα των Παραμέτρων της Παλινδρόμησης

- Σε επίπεδο σημαντικότητας $\alpha = 5\%$, οι ερμηνευτικές μεταβλητές οι οποίες θεωρούνται στατιστικά σημαντικές είναι αυτές των οποίων οι τιμή p – *value* είναι μικρότερη από 0,05. Στο παράδειγμα μας η b_0 και η b_2 .

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	332,6208172	37,52844531	8,86317	2,0738E-05	246,0800671	419,1615673	246,080067	419,1615673
Εμβαδόν σε τ.μ	0,46389432	0,204572939	2,26762	0,05308701	-0,007851724	0,935640364	-0,0078517	0,935640364
Παλιότητα σε έτη	-1,545381583	0,58440076	-2,64439	0,02951173	-2,893012153	-0,197751012	-2,8930122	-0,197751012

Έλεγχος για τη Σημαντικότητα των Παραμέτρων της Παλινδρόμησης

- Ειδικά στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης με μεγάλο πλήθος ανεξάρτητων μεταβλητών, τις ερμηνευτικές μεταβλητές οι οποίες δεν είναι στατιστικά σημαντικές δεν τις περιλαμβάνουμε στο γραμμικό μας μοντέλο.

Συντελεστές Συσχέτισης

- Μέσω της επιλογής

Data Analysis → Correlation

Του **Excel** μπορούμε να δούμε και του συντελεστές συσχέτισης όλων των μεταβλητών ανά δύο όπως φαίνεται στον παρακάτω πίνακα.

Συντελεστές Συσχέτισης – Παράδειγμα 1

	Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	Εμβαδόν σε τ.μ	Παλαιότη τα σε έτη
Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	1		
Εμβαδόν σε τ.μ	0,797233468	1	
Παλαιότητα σε έτη	-0,824963046	-0,635511594	1

Συντελεστές Συσχέτισης

- Σύμφωνα με τον διπλανό πίνακα, ο συντελεστής γραμμικής συσχέτισης των μεταβλητών X_1, Y είναι 0,7972. Οι μεταβλητές X_1, Y είναι θετικά συσχετισμένες. Δηλαδή αύξηση στο εμβαδόν του διαμερίσματος αναμένεται να αυξήσει την τιμή του.

	Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	Εμβαδόν σε τ.μ	Παλαιότητα σε έτη
Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	1		
Εμβαδόν σε τ.μ	0,797233468	1	
Παλαιότητα σε έτη	-0,824963046	-0,635511594	1

Συντελεστές Συσχέτισης

- Ο βαθμός της γραμμικής συσχέτισης είναι ισχυρή.

	Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	Εμβαδόν σε τ.μ	Παλαιότη τα σε έτη
Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	1		
Εμβαδόν σε τ.μ	0,797233468	1	
Παλαιότητα σε έτη	-0,824963046	-0,635511594	1

Συντελεστές Συσχέτισης

- Σύμφωνα με τον διπλανό πίνακα, ο συντελεστής γραμμικής συσχέτισης των μεταβλητών X_2, Y είναι $-0,82$. Οι μεταβλητές X_2, Y είναι αρνητικά συσχετισμένες. Δηλαδή αύξηση στην παλαιότητα του διαμερίσματος αναμένεται να μειώσει την τιμή του.

	Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	Εμβαδόν σε τ.μ	Παλαιότη τα σε έτη
Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	1		
Εμβαδόν σε τ.μ	0,797233468	1	
Παλαιότητα σε έτη	-0,824963046	-0,635511594	1

Συντελεστές Συσχέτισης

- Ο βαθμός της γραμμικής συσχέτισης είναι πολύ ισχυρή.

	Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	Εμβαδόν σε τ.μ	Παλαιότη τα σε έτη
Τιμή Πώλησης Διαμερίσματος σε χιλιάδες €	1		
Εμβαδόν σε τ.μ	0,797233468	1	
Παλαιότητα σε έτη	-0,824963046	-0,635511594	1

Συντελεστής Συσχέτισης - Υπενθυμίσεις

- Υπενθυμίζεται ότι:
 - Όταν ο συντελεστής συσχέτισης r είναι αρνητικός αριθμός τότε έχουμε αρνητική συσχέτιση.
 - Όταν ο συντελεστής συσχέτισης r είναι θετικός αριθμός τότε έχουμε θετική συσχέτιση.

Συντελεστής Συσχέτισης - Υπενθυμίσεις

- Ως προς τον βαθμός της γραμμικής συσχέτισης ισχύουν τα εξής:
 - Αν $r = \pm 1$ τότε **υπάρχει τέλεια γραμμική συσχέτιση.**
 - Αν $|r| \leq 0.3$ τότε λέμε ότι **δεν υπάρχει γραμμική συσχέτιση.**

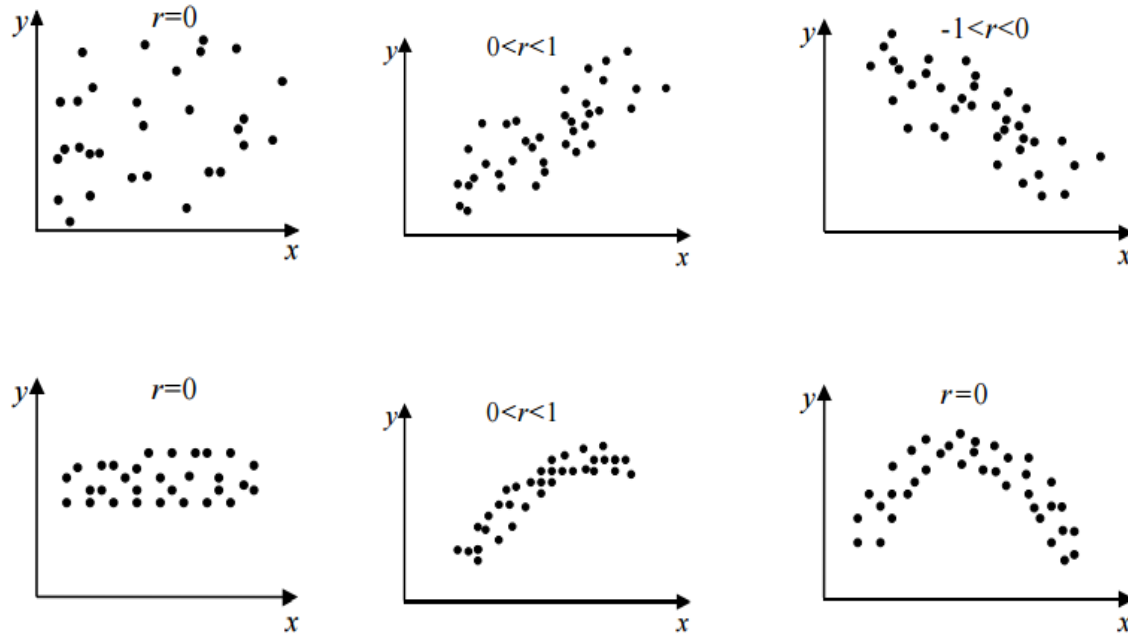
Συντελεστής Συσχέτισης - Υπενθυμίσεις

- Αν $0.3 < |r| \leq 0.5$ τότε υπάρχει ασθενής γραμμική συσχέτιση.
- Αν $0.5 < |r| \leq 0.7$ τότε υπάρχει μέση γραμμική συσχέτιση.
- Αν $0.7 < |r| \leq 0.8$ τότε υπάρχει ισχυρή γραμμική συσχέτιση.

Συντελεστής Συσχέτισης - Υπενθυμίσεις

- Αν $0.8 < |r| < 1$, τότε υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

Συντελεστής Συσχέτισης - Υπενθυμίσεις



73

Διάφορα είδη συσχέτισης

Μεικτή Παλινδρόμηση

- Σε κάποιες περιπτώσεις σε ένα μοντέλο παλινδρόμησης γίνεται χρήση μίας ή περισσότερων ποιοτικών μεταβλητών ως ερμηνευτικές μεταβλητές.

Μεικτή Παλινδρόμηση

- Για παράδειγμα, ένας ξενοδοχειακός όμιλος θα θέλει να γνωρίζει αν την υψηλή τουριστική περίοδο έχει περισσότερες κρατήσεις. Επομένως χρειάζεται την ποιοτική ερμηνευτική μεταβλητή «τουριστική περίοδος» με τιμές «χαμηλή» και «υψηλή».

Μεικτή Παλινδρόμηση

- Άλλο παράδειγμα ποιοτικής ερμηνευτικής μεταβλητής προκύπτει όταν μεσιτικό γραφείο ενδιαφέρεται να εκτιμήσει την τιμή των διαμερισμάτων σε σχέση με το αν είναι νοικιασμένα ή όχι. «Νοικιασμένο Διαμέρισμα», «Μη Νοικιασμένο Διαμέρισμα».

Μεικτή Παλινδρόμηση

- Στις περιπτώσεις που μία ποιοτική μεταβλητή έχει μόνο δύο τιμές, ένας εύκολος τρόπος να την ποσοτικοποιήσουμε αποτελεί η **δείκτρια μεταβλητή** ή η **ψευδομεταβλητή** (dummy variable).

Δείκτρια Μεταβλητή – Ψευδομεταβλητή

- Μία δείκτρια μεταβλητή είναι μία μεταβλητή που δείχνει κατά πόσο μία συνθήκη ισχύει.
- **Η δείκτρια μεταβλητή έχει:**
 - Την τιμή 1 όταν η συνθήκη ισχύει.
 - Την τιμή 0 όταν η συνθήκη δεν ισχύει.

Χρήση Δείκτριας Μεταβλητής στην Ανάλυση Παλινδρόμησης

- Η χρήση δείκτριων μεταβλητών στην ανάλυση παλινδρόμησης είναι πολύ απλή.
- Το μόνο που χρειάζεται είναι να κωδικοποιηθεί η ποιοτική δίτιμη μεταβλητή με την τιμή 1 όταν μία δεδομένη συνθήκη ισχύει και με την τιμή 0 όταν η ίδια συνθήκη δεν ισχύει.

Παράδειγμα 2

- Σύμφωνα με στοιχεία δημοσιευμένου σε εφημερίδα άρθρο, οι μεικτές αποδοχές των δημοσίων υπαλλήλων οι οποίοι είναι κάτοχοι πτυχίου ανώτατης εκπαίδευσης εξαρτώνται από την κατοχή ή όχι μεταπτυχιακού τίτλου σπουδών και από τα χρόνια προϋπηρεσίας τους.
- Στον παρακάτω πίνακα έχουμε καταγράψει τα στοιχεία από ένα τυχαίο δείγμα 12 δημοσίων υπαλλήλων, απόφοιτων ανώτατου εκπαιδευτικού ιδρύματος.

Παράδειγμα 2

Μεικτές Αποδοχές σε €	Προϋπηρεσία σε Έτη	Κατοχή Μεταπτυχιακού Τίτλου Σπουδών
902,72	7	ΟΧΙ
1018,78	15	ΟΧΙ
938,58	5	ΝΑΙ
1114,22	21	ΟΧΙ
884,52	1	ΝΑΙ
1087,65	15	ΝΑΙ
1147,63	13	ΟΧΙ
1222,71	27	ΟΧΙ
1152,87	19	ΝΑΙ
1256,69	25	ΝΑΙ
1056,8	7	ΝΑΙ
985,5	15	ΟΧΙ

Παράδειγμα 2

- Θα κωδικοποιήσουμε την δίτιμη ποιοτική μεταβλητή «**Κάτοχος Μεταπτυχιακού Τίτλου Σπουδών**» με την τιμή 1 όταν η απάντηση είναι *Ναι* και με την τιμή 0 όταν η απάντηση είναι *Όχι*.

Παράδειγμα 2

Μεικτές Αποδοχές σε €	Προϋπηρεσία σε Έτη	Κατοχή Μεταπτυχιακού Τίτλου Σπουδών
902,72	7	0
1018,78	15	0
938,58	5	1
1114,22	21	0
884,52	1	1
1087,65	15	1
1147,63	13	0
1222,71	27	0
1152,87	19	1
1256,69	25	1
1056,8	7	1
985,5	15	0

Παράδειγμα 2

- Για τη δημιουργία του μοντέλου πολλαπλής γραμμικής παλινδρόμησης ονομάζουμε τις μεταβλητές.

Έτσι έχουμε:

Παράδειγμα 2

- Y : Μεικτές Αποδοχές σε €.
- X_1 : Προϋπηρεσία σε Έτη.
- X_2 : Κατοχή Μεταπτυχιακού Τίτλου Σπουδών (ψευδομεταβλητή).

Παράδειγμα 2 – Υλοποίηση Μέσω **Excel**

- Παρακάτω βλέπουμε τον συγκεντρωτικό πίνακα που δίνει το **excel** για το μοντέλο της μεικτής γραμμικής παλινδρόμησης.

Παράδειγμα 2 – Υλοποίηση Μέσω Excel

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,908656663							
R Square	0,825656931							
Adjusted R Square	0,786914027							
Standard Error	56,1066493							
Observations	12							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	134173,3062	67086,7	21,31117813	0,000385763			
Residual	9	28331,60487	3147,96					
Total	11	162504,9111						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	832,2754044	42,40727939	19,6258	1,07248E-08	736,3434736	928,2073353	736,3434736	928,2073353
Προμηθευσία σε Έτη	14,26436299	2,185052203	6,52816	0,000107866	9,321431501	19,20729448	9,321431501	19,20729448
Κατοχή Μεταπτυχιακού Τίτλου Σπουδών	59,4039063	33,74866395	1,76019	0,112229461	-16,94087559	135,7486882	-16,94087559	135,7486882

Παράδειγμα 2 – Υλοποίηση Μέσω Excel

Από το πρώτο πλαίσιο του πίνακα βλέπουμε ότι ο συντελεστής προσδιορισμού (***R Square***) είναι

$$R^2 = 82565 = 82,565\%$$

Ενώ ο προσαρμοσμένος (adjusted) συντελεστής πολλαπλού προσδιορισμού (***Adjusted R Square***)

$$\bar{R}^2 = 0,7869 = 78,69\%$$

Regression Statistics	
Multiple R	0,908656663
R Square	0,825656931
Adjusted R Square	0,786914027
Standard Error	56,1066493
Observations	12

Παράδειγμα 2 – Υλοποίηση Μέσω Excel

- Δηλαδή το **78,69%** της μεταβλητότητας των μεικτών αποδοχών των δημοσίων υπαλλήλων, ερμηνεύεται από κοινού από την προϋπηρεσία τους σε έτη και από την κατοχή τίτλου μεταπτυχιακών σπουδών.

Παράδειγμα 2 – Υλοποίηση Μέσω Excel

- Το τυπικό σφάλμα εκτίμησης είναι (Standard Error) είναι $S_E = 56,1$.

<i>Regression Statistics</i>	
Multiple R	0,908656663
R Square	0,825656931
Adjusted R Square	0,786914027
Standard Error	56,1066493
Observations	12

Παράδειγμα 2 – Υλοποίηση Μέσω Excel

- Το τελευταίο πλαίσιο του πίνακα μας δίνει την εξίσωση της παλινδρόμησης που είναι

$$Y = 832,27 + 14,26X_1 + 59,4X_2 + \varepsilon$$

Δηλαδή,

$$\hat{Y}_i = 832,27 + 14,26X_{1,i} + 59,4X_{2,i}$$

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95,0%</i>	<i>Upper 95,0%</i>
Intercept	832,2754044	42,40727939	19,6258	1,07248E-08	736,3434736	928,2073353	736,3434736	928,2073353
Προϋπηρεσία σε Έτη	14,26436299	2,185052203	6,52816	0,000107866	9,321431501	19,20729448	9,321431501	19,20729448
Κατοχή Μεταπτυχιακού Τίτλου Σπουδών	59,4039063	33,74866395	1,76019	0,112229461	-16,94087559	135,7486882	-16,94087559	135,7486882

Παράδειγμα 2

Ερμηνεία της σταθεράς $b_0 = 832,27$:

- Δεδομένου ότι ένας δημόσιος υπάλληλος έχει μηδενική προϋπηρεσία και δεν είναι κάτοχος μεταπτυχιακού τίτλου σπουδών, τότε οι αναμενόμενες μεικτές του αποδοχές είναι **832,27€**.

Παράδειγμα 2

- Ο συντελεστής μερικής κλίσης $b_1 = 14,26$ μας λέει ότι για κάθε ένα έτος παραπάνω προϋπηρεσίας και δεδομένου ότι το προφίλ του δημοσίου υπαλλήλου δεν αλλάζει ως προς την κατοχή μεταπτυχιακού τίτλου σπουδών, οι μεικτές του αποδοχές αναμένεται να αυξάνονται κατά **14,26€**.

Παράδειγμα 2

- Ο συντελεστής μερικής κλίσης $b_2 = 59,4$ της ψευδομεταβλητής, μας λέει ότι όταν ένας δημόσιος υπάλληλος είναι κάτοχος μεταπτυχιακού τίτλου σπουδών οι μεικτές του αποδοχές αυξάνονται κατά **59,4€**.
- Όταν ο δημόσιος υπάλληλος δεν είναι κάτοχος μεταπτυχιακού τίτλου σπουδών η παραπάνω αύξηση δεν ισχύει.

Παράδειγμα 2

- Όπως αναφέραμε και στο προηγούμενο παράδειγμα, σε επίπεδο σημαντικότητας $\alpha = 5\%$, οι ερμηνευτικές μεταβλητές οι οποίες θεωρούνται στατιστικά σημαντικές είναι αυτές των οποίων οι τιμή $p - value$ είναι μικρότερη από 0,05. Στο παράδειγμα μας η b_0 και η b_1 .

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	832,2754044	42,40727939	19,6258	1,07248E-08	736,3434736	928,2073353	736,3434736	928,2073353
Προπηρεσία σε Έτη	14,26436299	2,185052203	6,52816	0,000107866	9,321431501	19,20729448	9,321431501	19,20729448
Κατοχή Μεταπτυχιακού Τίτλου Σπουδών	59,4039063	33,74866395	1,76019	0,112229461	-16,94087559	135,7486882	-16,94087559	135,7486882

Βιβλιογραφία

Am. Aczel, J. Sounderprandian, Στατιστική σκέψη στον κόσμο των επιχειρήσεων, Broken Hill, 2013

ΣΤ. Δημητριάδης, Στατιστική Επιχειρήσεων με Εφαρμογές σε SPSS και LISREL, Εκδόσεις Κριτική, 2016

Γ. Κ. Χρήστου, Εισαγωγή στην Οικονομετρία, Α' Τόμος Εκδόσεις Gutenberg, 2007

Χαλικιάς Μ. Ποσοτική Ανάλυση και Στοιχεία Θεωρίας Αποφάσεων στη Διοίκηση και Οικονομία με Χρήση Λογισμικών EXCEL, ISALOS και SPSS, Εκδόσεις Broken Hill Publishers Ltd, 2021

Alder, H. L. and Roessler, E. B. Introduction to Probability and Statistics. 6th edition, W. H. Freeman & Company.