

**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΕΦΑΡΜΟΣΜΕΝΗΣ ΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ»

ΚΑΤΕΥΘΥΝΣΗ

«ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ»

Master of Science in

Business Administration, Analytics and Information Systems

**Data Science (R)**

**Assignment 1**

**Κωνσταντίνος Κουτσομπίνας**

Αθήνα 29/05/2025

Περιεχόμενα

[Άσκηση 1 3](#_Toc200052230)

[I. Βασικά Χαρακτηριστικά 3](#_Toc200052231)

[II. Pair Plots 3](#_Toc200052232)

[III. Ανάλυση ιστογραμμάτων 5](#_Toc200052233)

[IV. Σύνορα με ποταμό 5](#_Toc200052234)

[V. Διάμεση αναλογία μαθητών και δασκάλων 5](#_Toc200052235)

[VI. Προάστια με μικρότερη αξία κατοικιών 5](#_Toc200052236)

[VII. Προάστια με μέσο όρο δωματίων ανά σπίτι > 8 6](#_Toc200052237)

[Άσκηση 2 7](#_Toc200052238)

[I. Γραμμική παλινδρόμηση με όλες τις μεταβλητές 7](#_Toc200052239)

[II. Σύνοψη αποτελεσμάτων 7](#_Toc200052240)

[III. Διαστήματα εμπιστοσύνης 8](#_Toc200052241)

[IV. Πρόβλεψη 8](#_Toc200052242)

[V. Backward Stepwise Regression 9](#_Toc200052243)

[VI. Train Test Split 9](#_Toc200052244)

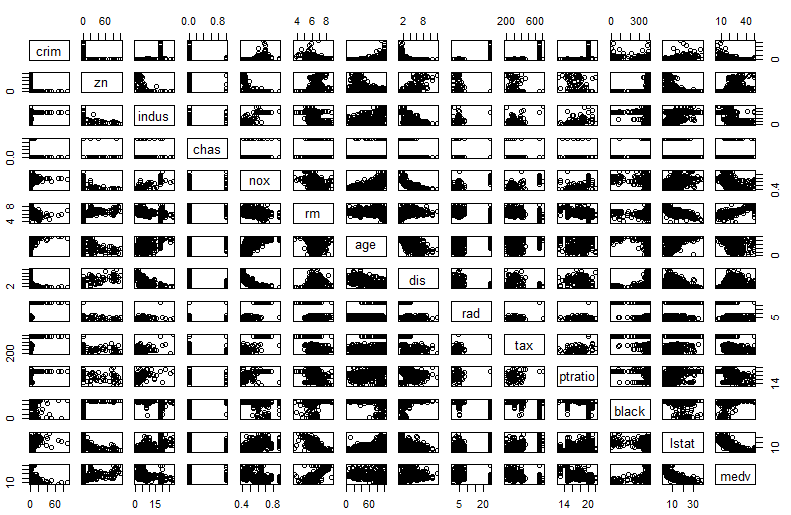
# Άσκηση 1

## Βασικά Χαρακτηριστικά

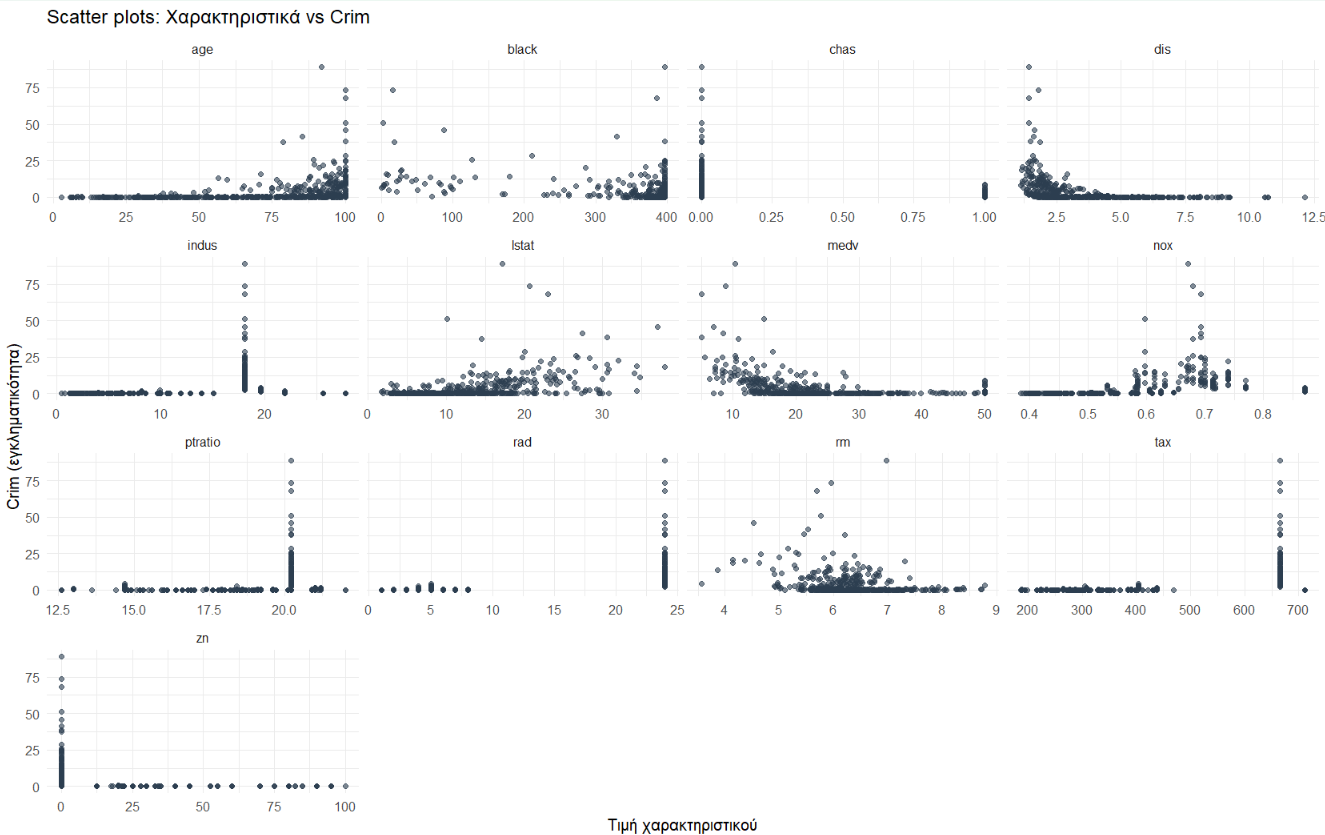
To Boston dataset περιέχει 506 γραμμές και 14 στήλες. Δηλαδή περιέχει 14 γνωρίσματα για 506 προάστια της Βοστώνης. Η περιγραφή των ορισμάτων είναι η εξής (όπως δίνεται από το ?Boston):

* Crim: per capita crime rate by town.
* Zn: proportion of residential land zoned for lots over 25,000 sq.ft.
* Indus: proportion of non-retail business acres per town.
* Chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
* Nox: nitrogen oxides concentration (parts per 10 million).
* Rm: average number of rooms per dwelling.
* Age: proportion of owner-occupied units built prior to 1940.
* Dis: weighted mean of distances to five Boston employment centres.
* Rad: index of accessibility to radial highways.
* Tax: full-value property-tax rate per $10,000.
* Ptratio: pupil-teacher ratio by town.
* Black: 1000(Bk−0.63)21000(*Bk*−0.63)2 where Bk*Bk* is the proportion of blacks by town.
* Lstat: lower status of the population (percent).
* Medv: median value of owner-occupied homes in $1000s.

## Pair Plots

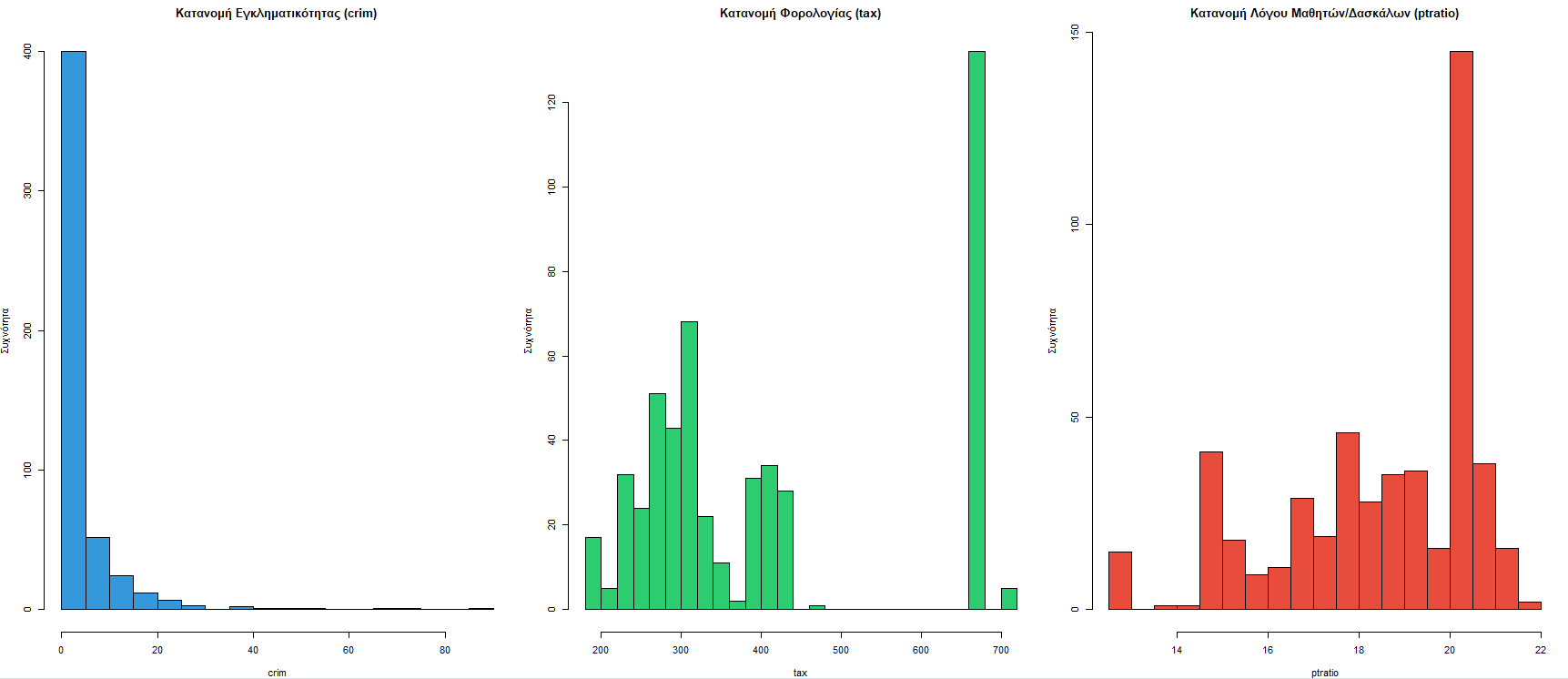


Ας παρατηρήσουμε καλύτερα τις συσχετίσεις κάθε μεταβλητής με τον δείκτη crim.



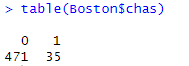
* Age: Στις παλαιότερες γειτονιές σημειώνεται μεγαλύτερη μέση εγκληματικότητα.
* Black: Δεν υπάρχει κάποια ξεκάθαρη σχέση, ίσως παρατηρούμε αύξηση εγκληματικότητας στα 2 άκρα αυτού του δείκτη.
* Chas: τα προάστια που δεν συνορεύουν με τον ποταμό έχουν υψηλότερο μέσο δείκτη εγκληματικότητας
* Dis: Η μεγαλύτερη απόσταση με τα κέντρα εργασίας δείχνει να μειώνει τη μέση εγκληματικότητα, (ίσως και λόγω μικρότερου πληθυσμού).
* Indus: Τα περισσότερα προάστεια με υψηλή εγκληματικότητα έχουν δείκτη indus κοντά στο 18.
* Lstat: Οι περισσότερες περιοχές με χαμηλό δείκτη lstat έχουν και χαμηλή μέση εγκληματικότητα, όμως φαίνεται να υπάρχουν και εξαιρέσεις.
* Medv: Τα προάστια με τα φθηνότερα διαμερίσματα κατά μέσο όρο δείχνουν να έχουν υψηλότερη εγκληματικότητα (παρατηρούμε όμως και μία μικρή αύξηση στις ακριβότερες περιοχές)
* Nox: Όσο μεγαλώνει η ρύπανση μεγαλώνει και η εγκληματικότητα, μέχρι που ξαναπέφτει (πιθανώς διότι μετά από κάποιο επίπεδο μικραίνει σημαντικά ο πληθυσμός του προαστίου).
* Ptratio: Οι περιοχές με την μεγαλύτερη εγκληματικότητα έχουν δείκτη ptratio κοντά στο 20.
* Rad: Οι υψηλές τιμές εγκληματικότητας δείχνουν να φανερώνονται σε προάστια με μεγάλο δείκτη προσβασιμότητας σε κεντρικούς δρόμους.
* Rm: Στα προάστια με υψηλότερο μέσο όρο αριθμού δωματίων ανά σπίτι, παρουσιάζεται χαμηλότερη μέση εγκληματικότητα (αν και φαίνεται να υπάρχουν εξαιρέσεις)
* Tax: Οι περισσότεροι δείκτες υψηλής εγκληματικότητας βρίσκονται σε προάστια με υψηλή φορολογία.
* Zn: τα περισσότερα προάστια με zn > 0 έχουν εγκληματικότητα κοντά στο 0.

## Ανάλυση ιστογραμμάτων



* Ο δείκτης εγκληματικότητας δείχνει να είναι πολύ λοξός προς τα δεξιά. Δηλαδή το μεγαλύτερο μέρος των προαστίων της Βοστώνης έχει χαμηλή εγκληματικότητα, αλλά υπάρχουν περιοχές με πολύ μεγαλύτερα ποσοστά από το μέσο, όρο.
* Ο δείκτης φορολογίας, δείχνει να χωρίζεται στα 2, σε μία κατανομή με σχετικά χαμηλή φορολογία, αλλά υπάρχει και μία πολύ μεγάλη μερίδα των προαστίων που έχει πολύ υψηλή φορολογία.
* Ο λόγος μαθητών δασκάλων έχει κατανομή λοξή προς τα αριστερά. Δηλαδή το μεγαλύτερο ποσοστό των προαστίων έχει υψηλή τιμή αυτού του λόγου, αλλά υπάρχουν αρκετές περιοχές με σημαντικά μικρότερες αναλογίες.

## Σύνορα με ποταμό



Βλέπουμε ότι 471/506 (93%) προάστια δεν συνορεύουν με τον ποταμό, ενώ 35/506 (7%) συνορεύουν.

## Διάμεση αναλογία μαθητών και δασκάλων



Η διάμεση αναλογία μαθητών και δασκάλων είναι 18.45553 μαθητές ανά δάσκαλο.

## Προάστια με μικρότερη αξία κατοικιών

A close up of a text

AI-generated content may be incorrect.

Βλέπουμε ότι τα προάστια (2) με τη μικρότερη διάμεση αξία κατοικιών σε σχέση με τα υπόλοιπα έχουν :

* Πολύ υψηλότερη μέση εγκληματικότητα
* Μηδενικό δείκτη zn ενώ τα άλλα προάστια έχουν μ.ο. 11.41
* Σχετικά μεγάλη βιομηχανική περιοχή
* Κανένα δεν συνορεύει με τον ποταμό
* Αρκετά υψηλή ρύπανση
* Σχετικά χαμηλότερο αριθμό δωματίων
* Είναι οι παλαιότερες περιοχές
* Μικρότερη απόσταση από τα κέντρα εργασίας
* Μεγαλύτερη προσβασιμότητα σε κεντρικούς δρόμους
* Είναι από τις πιο υψηλά φορολογούμενες
* Σχετικά μεγάλη αναλογία μαθητών / καθηγητών
* Είναι κοντά στο διάμεσο του δείκτη black
* Υψηλή τιμή δείκτη low status

## Προάστια με μέσο όρο δωματίων ανά σπίτι > 8

A close up of a text

AI-generated content may be incorrect.

Τα προάστια με μέσο όρο δωματίων ανά σπίτι > 8 έχουν:

* Χαμηλότερη εγκληματικότητα
* Κατανομή δείκτη zn παρόμοια με τα υπόλοιπα
* Σχετικά χαμηλότερη βιομηχανική περιοχή
* Σχετικά υψηλότερο ποσοστό περιοχών που συνορεύουν με τον ποταμό
* Ελάχιστα χαμηλότερη μέση ρύπανση
* Μέτρια παλιές περιοχές
* Σχετικά κοντά στις κέντρα εργασίας
* Λίγο χαμηλότερο δείκτη προσβασιμότητας σε μεγάλους δρόμους
* Ελαφρώς χαμηλότερη φορολογία
* Ελάχιστα χαμηλότερη μέση αναλογία μαθητών/δασκάλων
* Είναι κοντά στο διάμεσο του δείκτη black
* Αρκετά χαμηλό δείκτη low status
* Υψηλότερη διάμεση τιμή κατοικιών

# Άσκηση 2

## Γραμμική παλινδρόμηση με όλες τις μεταβλητές

Καλούμε την

1. fitAll <- lm(Sales~., Carseats)

## Σύνοψη αποτελεσμάτων

A screenshot of a computer

AI-generated content may be incorrect.

Παρατηρήσεις:

Α) Ερμηνεία των συντελεστών:

* Η σταθερά έχει τιμή 5.66. Δηλαδή αν όλα τα άλλα είναι 0 οι πωλήσεις είναι 5.66 (δύσκολη πραγματική ερμηνεία).
* Η τιμή του ανταγωνισμού έχει συντελεστή 0.093 δηλαδή, όσο αυξάνεται η τιμή των ανταγωνιστών, αυξάνονται και οι πωλήσεις του καταστήματος, όπως και θα αναμέναμε.
* Το εισόδημα των καταναλωτών έχει συντελεστή 0.016 δηλαδή επηρεάζει θετικά τις πωλήσεις.
* Η διαφήμιση έχει επίσης θετικό συντελεστή, 0.123, επομένως και αυτή επηρεάζει θετικά της πωλήσεις
* Ο πληθυσμός έχει συντελεστή 0.0002, st error 0.011 και pvalue 0.575 που υποδεικνύει **μη σημαντικότητα**
* Η τιμή έχει συντελεστή -0.095, δηλαδή αύξηση της τιμής μειώνει της πωλήσεις, και το αντίστροφο.
* Οι dummy μεταβλητές ShelveLocGood και ShelveLocMedium έχουν συντελεστές 4.85 και 1.96 αντίστοιχα, που σημαίνει ότι όσο καλύτερη είναι η ποιότητα της θέσης στο ράφι, τόσο περισσότερες είναι οι πωλήσεις
* Η μέση ηλικία του τοπικού πληθυσμού έχει συντελεστή -0.046. Πιθανώς όσο αυξάνεται αυτός ο δείκτης, μειώνεται το πλήθος παιδιών στον τοπικό πληθυσμό, επομένως και η ανάγκη για παιδικά καρεκλάκια
* Οι 3 τελευταίοι συντελεστές (Education, UrbanYes, USYes) έχουν μεγάλα pvalues, και το st error τους είναι συγκριτικά μεγάλο σε σχέση με τη τιμή των συντελεστών τους. Συνεπώς **δεν είναι στατιστικά σημαντικοί και δεν χρίζουν ερμηνεία.**

Β) Η συνολική στατιστική σημαντικότητα είναι αρκετά μεγάλη. Βλέπουμε pvalue < το οποίο εκφράζει τη πιθανότητα, καμία μεταβλητή να μην είναι στατιστικά σημαντική στην εξήγηση των πωλήσεων

Γ) Έχουμε , που υποδεικνύει πολύ ισχυρή σχέση μεταξύ των επεξηγηματικών μεταβλητών και των πωλήσεων, καθώς εξηγείται το 87% της διακύμανσης των πωλήσεων.

Δ) Οι μεταβλητές CompPrice, Income, Adverticing, Price, ShelveLoc και Age έχουν μικρά pvalues και συνεπώς είναι στατιστικά σημαντικές. Ενώ οι μεταβλητές Population, Education, Urban και US δεν είναι στατιστικά σημαντικές.

Ε) Για τις στατιστικά σημαντικές μεταβλητές τα πρόσημα είναι αναμενόμενα και ερμηνεύονται ως εξής:

|  |  |  |
| --- | --- | --- |
| Μεταβλητή | Πρόσημο | Εξήγηση |
| CompPrice | + | Όταν ο ανταγωνιστής είναι ακριβός, οι πωλήσεις μας αυξάνονται |
| Advertising | + | Η διαφήμιση αυξάνει τις πωλήσεις |
| ShelveLoc | + | Καλύτερη θέση στο ράφι -> αυξημένες πωλήσεις |
| Income | + | Μεγαλύτερο εισόδημα -> αυξημένες πωλήσεις |
| Price | - | Αυξημένη τιμή -> μειωμένες πωλήσεις |
| Age | - | Μεγαλύτερο κοινό -> μικρότερη ανάγκη για παιδικά καρεκλάκια |

## Διαστήματα εμπιστοσύνης

A screenshot of a computer

AI-generated content may be incorrect.

* Το διάστημα εμπιστοσύνης κάθε συντελεστή, εκφράζει ένα διάστημα για το οποίο είμαστε κατά 95% σίγουροι, πως μέσα σε αυτό βρίσκεται η πραγματική τιμή του συντελεστή.
* Παρατηρούμε ότι για κάποιες μεταβλητές (Population, Education, Urban, Us) το διάστημα αυτό είναι πολύ ευρύ, και δεν διατηρεί σταθερό πρόσημο. Ενώ για τις υπόλοιπες είναι σχετικά μικρό και διατηρεί σταθερό πρόσημο. Αυτό υποδεικνύει ότι οι μεταβλητές με μεγάλο διάστημα εμπιστοσύνης δεν είναι στατιστικά σημαντικές, και συνεπώς δεν μπορούμε να είμαστε ιδιαίτερα σίγουροι για την τιμή τους. Επιπλέον παρατηρούμε ότι οι μεταβλητές με μεγάλο διάστημα εμπιστοσύνης είναι οι ίδιες με αυτές που υποδείξαμε ως μη σημαντικές στο προηγούμενο σκέλος, με βάση τις τιμές του pvalue.

## Πρόβλεψη

A computer screen shot of a computer code

AI-generated content may be incorrect.

* Παρατηρούμε ότι η πρόβλεψη για αυτό το νέο Χ που κατασκευάσαμε από τις μέσες τιμές των αριθμητικών μεταβλητών και τις επικρατούσες τιμές των κατηγορικών μεταβλητών, είναι 7.321973.
* Το διάστημα εμπιστοσύνης είναι σχετικά στενό (7.13313, 7.510817) ενώ το διάστημα πρόβλεψης είναι σημαντικά ευρύτερο (5.309765, 9.334181). Αυτό συμβαίνει διότι το διάστημα εμπιστοσύνης εκφράζει ένα διάστημα **μέσα στο οποίο αναμένουμε να βρίσκεται η μέση τιμή της εξαρτημένης μεταβλητής** για τις δοσμένες τιμές των ανεξάρτητων μεταβλητών, με ένα συγκεκριμένο επίπεδο εμπιστοσύνης (π.χ. 95%). Αντίθετα, το **διάστημα πρόβλεψης** περιλαμβάνει όχι μόνο την αβεβαιότητα γύρω από τη μέση τιμή, αλλά και τη **φυσική μεταβλητότητα των μεμονωμένων παρατηρήσεων** — δηλαδή πόσο μπορεί να διαφέρει μία νέα, ατομική παρατήρηση από τη μέση τιμή.

## Backward Stepwise Regression

1. step <- stepAIC(fitAll, direction="backward")

2. summary(step)

A screenshot of a computer

AI-generated content may be incorrect.

Παραπάνω βλέπουμε τα αποτελέσματα του backward stepwise regression. Παρατηρούμε ότι ο αλγόριθμος αφαιρεί με τη σειρά τις μεταβλητές **population, urban, education, us**, καθώς δεν είναι στατιστικά σημαντικές. Παρατηρούμε επίσης ότι αυτές οι μεταβλητές είναι οι ίδιες που χαρακτηρίσαμε ως μη σημαντικές και στα ερωτήματα [II](#_Σύνοψη_αποτελεσμάτων) και [III](#_Διαστήματα_εμπιστοσύνης).

## Train Test Split

1. # vi) partition train test

2. set.seed(123)

3. training.samples <- createDataPartition(Carseats$Sales, p = 0.7, list = FALSE)

4. train.data <- Carseats[training.samples, ]

5. test.data <- Carseats[-training.samples, ]

6.

7. #use only the significant values from previous step

8. model <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age, data = train.data)

9. summary(model)

10.

11. # Predict in test data

12. predictions <- predict(model, newdata = test.data)

13.

14. # r2 of test data

15. R2(predictions, test.data$Sales)

16.

17. # 5-fold

18. control <- trainControl(method = "cv", number = 5)

19.

20. cv\_model <- train(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age,

21. data = Carseats, method = "lm",

22. trControl = control)

Παίρνουμε τα εξής αποτελέσματα :

= 0.876

= 0.8605736

= 0.8675214

* Παρατηρούμε λοιπόν ότι τα είναι αρκετά υψηλά ( της διακύμανσης των Sales εξηγείται από τις μεταβλητές) και κοντά μεταξύ τους.
* Στο train set είναι ελαφρώς καλύτερο, ωστόσο δεν υπάρχει overfit καθώς η διαφορά είναι ελάχιστη. Αυτό είναι αναμενόμενο καθώς έχουμε ένα μοντέλο το οποίο έχει μόνο στατιστικά σημαντικές μεταβλητές.
* Επίσης παρατηρούμε ότι το cross validation δίνει αποτελέσματα πολύ κοντά στο αρχικό split. Συνεπώς το μοντέλο παρουσιάζει σταθερή και αξιόπιστη απόδοση.