# Υπολογιστική Στατιστική

Κατερίνα Ορφανογιαννάκη

Τμήμα Μαθηματικών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
korfanog@math.uoa.gr

2020-2021

# Μέθοδος Cross-Validation Εισαγωγή

Η μέθοδος Cross-Validation χρησιμοποιείται τόσο για τη μελέτη της (καλής) προσαρμογής ενός μοντέλου στα δεδομένα όσο και για τη σύγκριση ανάμεσα σε μοντέλα.
Συγκεκριμένα η μέθοδος Cross-Validation μας επιτρέπει:

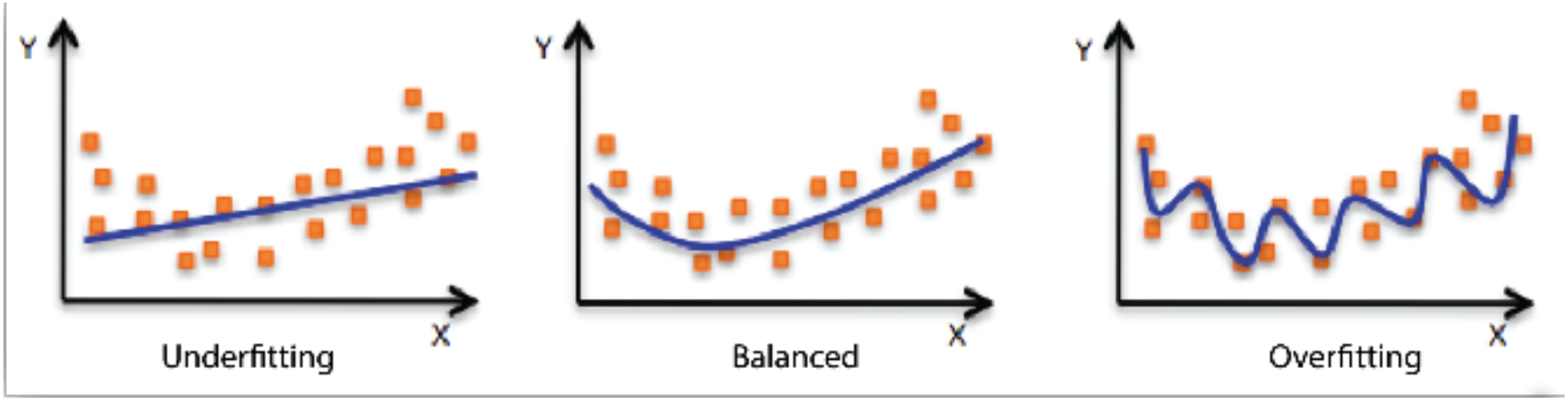
1. Να εξετάσουμε την καλή προσαρμογή ενός στατιστικού μοντέλου στα δεδομένα μας.
2. Να διαλέξουμε ανάμεσα σε μοντέλα.

Με τη μέθοδο Cross-Validation ΔΕΝ εξετάζουμε την ποιότητα εκτιμητριών.

# Γιατί Cross-Validation

- Χρειαζόμαστε τη μέθοδο Cross-Validation για να εξετάσουμε την προσαρμογή ενός μοντέλου στα δεδομένα αποφεύγοντας περιπτώσεις υπερμοντελοποίησης.
- Definition overfitting: "The production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably."

# Παράδειγμα



# Overfitting

- Περιπτώσεις υπερμοντελοποίησης έχουμε όταν το μοντέλο προσαρμόζεται υπερβολικά καλά στα δεδομένα. Σε αυτές τις περιπτώσεις το μοντέλο είναι υπερβολικά περίπλοκο και έχει περισσότερες παραμέτρους από όσους χρειάζονται παραγματικά.
- Σε περιπτώσεις φωλιασμένων μοντέλων γνωρίζουμε ότι όσο περισσότερες παραμέτρους προσθέτουμε στο μοντέλο τόσο περισσότερο καλά το μοντέλο προσαρμόζεται στα δεδομένα. Χρειάζονται όμως όλες οι παράμετροι;

# Σύνηθες πρακτική για την καλή προσαρμογή

- Συνήθως προσαρμόζουμε το μοντέλο στα δεδομένα μας και μετά χρησιμοποιώντας κάποιο κατάλληλο μέτρο εξετάζουμε την καλή προσαρμογή του.
- Κλασσικό παράδειγμα στην Γραμμική Παλινδρόμηση είναι ο συντελεστής προσδιορισμού. Μειονέκτημα: Χρησιμοποιούμε τα ίδια δεδομένα δύο φορές. Μία για να προσαρμόσουμε το μοντέλο δεύτερη και για να δούμε πόσο καλό είναι.

# Η μέθοδος

Έστω δύο μοντέλα $g$ και $h$ και θέλουμε να εξετάσουμε ποιο από τα 2 μοντέλα προσαρμόζεται καλύτερα στις $n$ παρατηρήσεις μας $x_1, \ldots, x_n$.
Αφήνουμε μια παρατήρηση έξω κάθε φορά,εκτιμάμε τις παραμέτρους κάθε μοντέλου χρησιμοποιώντας μόνο τις υπόλοιπες παρατηρήσεις, και στη συνέχεια, με βάση τις εκτιμήσεις που έχουμε πάρει για τις παραμέτρους του μοντέλου, προβλέπουμε την τιμή που αφήσαμε έξω.
Συμβολίζουμε με $\hat{g}_{-i}$ το μοντέλο που έχει εκτιμηθεί χωρίς την παρατήρηση $i$.
Έστω $\hat{g}_{-i}(x_i)$και $\hat{h}_{-i}(x_i)$ οι προβλέψεις για την τιμή $x_i$ από τα μοντέλα $\hat{g}_{-i}$ και $\hat{h}_{-i}$ αντίστοιχα.
Οι τιμές $y_i - \hat{g}_{-i}(x_i)$ και $y_i - \hat{h}_{-i}(x_i)$ είναι τα σφάλματα της πρόβλεψης για κάθε μοντέλο.

# PRESS

Έστω η συνάρτηση

$$CV(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}_{-i}(x_i))^2.$$

Η ποσότητα $nCV(g)$ αναφέρεται ως άθροισμα τετραγώνων του προβλεπτικού σφάλματος (prediction error sum of squares, PRESS) για το μοντέλο $g$. Συγκρίνοντας τις τιμές της συνάρτησης $CV$ για κάθε μοντέλο μπορεί κανείς να δει ποιο μοντέλο είναι καλύτερο. Έτσι αν $CV(g) < CV(h)$ τότε το μοντέλο $g$ είναι καλύτερο από το $h$ καθώς έχει μικρότερο PRESS.

# Σχόλια:

- Η μέθοδος Cross-validation απαιτεί πολλούς υπολογισμούς, αν και υπάρχουν σε πολλές περιπτώσεις ικανοποιητικοί αλγόριθμοι για να μειώσουν τον όγκο των υπολογισμών που απαιτούνται.
- Η μέθοδος Cross-validation δεν επηρεάζεται από ακραίες τιμές με μεγάλη επίδραση που δεν οφείλονται στο μοντέλο και οι οποίες μπορούν να καθορίζουν τις εκτιμήτριες του μοντέλου.
- Το κριτήριο PRESS και μπορεί να χρησιμοποιηθεί για την σύγκριση φωλιασμένων μοντέλων.
- Μπορούμε να χρησιμοποιήσουμε διάφορες άλλες προσεγγίσεις για να μετρήσουμε την καλή προσαρμογή του μοντέλου όπως π.χ. να χρησιμοποιήσουμε απόλυτες τιμές αντί να υψώσουμε στο τετράγωνο.
- Το κριτήριο PRESS εκτός από την καλή προσαρμογή εξετάζει και την προβλεπτική ικανότητα του μοντέλου.

# Παράδειγμα: Γραμμική Παλινδρόμηση

Γνωρίζουμε πως γενικά το γραμμικό μοντέλο με τη χρήση πινάκων μπορεί να γραφτεί ως εξής

$$\mathbf{Y} = \mathbf{X}\beta$$

όπου **Y** είναι ένα $n \times 1$ διάνυσμα με τις τιμές της εξαρτημένης μεταβλητής, **X** είναι ένας $n \times (p + 1)$ πίνακας σχεδιασμού με τις $p$ ανεξάρτητες μεταβλητές (και συνήθως την πρώτη στήλη όλο μονάδες για να αναπαριστούν τη σταθερά) και $\beta$ είναι ένα $(p + 1) \times 1$ διάνυσμα με τους συντελεστές.

# Παράδειγμα: Γραμμική Παλινδρόμηση (συνέχεια)

Η εκτιμήτρια ελαχίστων τετραγώνων είναι η

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}.$$

Το διάνυσμα με τις προβλέψεις $\hat{\mathbf{Y}}$ δίνεται από το:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}\mathbf{Y}.$$

όπου $\mathbf{A} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ είναι ένας πίνακας που ονομάζεται $hat - matrix$ και έχει μερικές σημαντικές ιδιότητες. Αν συμβολίσουμε με $a_{ij}$ το $ij$ στοιχείο του πίνακα $A$ τότε αποδεικνύεται εύκολα η σχέση

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - a_{ii}}$$

και επομένως χρειάζεται να γίνει προσαρμογή του μοντέλου μόνο μία φορά και να αποθηκευθεί ο πίνακας $A$. Τα διαγώνια στοιχεία του πίνακα $A$ ($a_{ii}$) συνήθως αναφέρονται στη βιβλιογραφία ως *leverages* και χρησιμοποιούνται για να μετρήσουμε αν μια παρατήρηση μπορεί να θεωρηθεί πως έχει μεγάλη επίδραση στην εκτίμηση του μοντέλου.

# Παράδειγμα χρήσης της μεθόδου

Τα δεδομένα που υπάρχουν στον επόμενο πίνακα αφορούν 18 παρατηρήσεις όπου για μια συγκεκριμένη σοδιά καλαμποκιού χρησιμοποιήθηκαν λιπάσματα με φώσφορο τόσο σε οργανική όσο και σε ανόργανη μορφή. Σκοπός του πειράματος ήταν να μελετηθεί κατά πόσο ο φώσφορος πέρασε στην παραγωγή του καλαμποκιού. Έτσι τα δεδομένα αποτελούνται από 3 μεταβλητές

$Y$ = Ποσότητα φωσφόρου στην παραγωγή καλαμποκιού

$X_1$ = Ποσότητα ανόργανου φωσφόρου στο λίπασμα

$X_2$= Ποσότητα οργανικού φωσφόρου στο λίπασμα

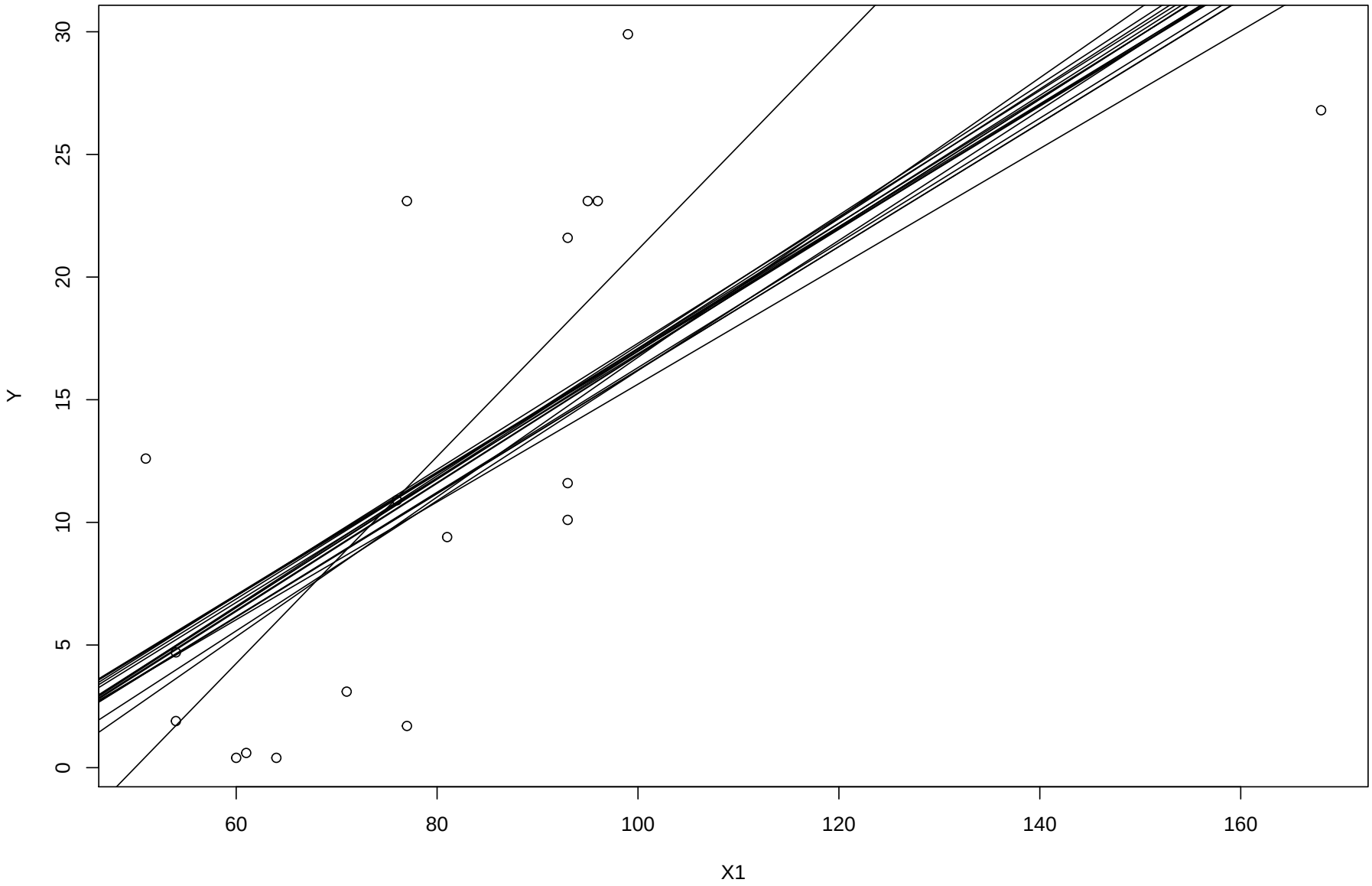
Σκοπός της μελέτης ήταν να μελετηθεί η σχέση ανάμεσα στο φώσφορο του λιπάσματος και της παραγωγής προσαρμόζοντας κατάλληλα γραμμικά μοντέλα.

# Δεδομένα

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 0.4 | 0.4 | 3.1 | 0.6 | 4.7 | 1.7 | 9.4 | 10.1 | 11.6 |
| $X_1$ | 64 | 60 | 71 | 61 | 54 | 77 | 81 | 93 | 93 |
| $X_2$ | 53 | 23 | 19 | 34 | 24 | 65 | 44 | 31 | 29 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 12.6 | 10.9 | 23.1 | 23.1 | 21.6 | 23.1 | 1.9 | 26.8 | 29.9 |
| $X_1$ | 51 | 76 | 96 | 77 | 93 | 95 | 54 | 168 | 99 |
| $X_2$ | 58 | 37 | 46 | 50 | 44 | 56 | 36 | 58 | 51 |

# Διάγραμμα σημείων και ευθείες



# Κριτήρια επιλογής μοντέλου

| Επεξηγηματικές Μεταβλητές | PRESS | Συντελεστής Προσδιορισμού |
|---|---|---|
| $X_1$ | 1526,6 | 48% |
| $X_2$ | 1809,10 | 21% |
| $X_1 + X_2$ | 1601,55 | 53% |
| $X_1 + X_1^2$ | 26248,5 | 55% |
| $X_1 + X_1^2 + X_1^3$ | 19676,2 | 68% |

# Συμπεράσματα

- Το κριτήριο PRESS δεν έχει τη μονότονη συμπεριφορά του συντελεστή προσδιορισμού και επομένως εισάγοντας μια καινούρια επεξηγηματική μεταβλητή στις ήδη υπάρχουσες δεν βελτιώνεται αναγκαστικά.
- Το κριτήριο μπορεί να χρησιμοποιηθεί ανάμεσα και σε φωλιασμένα μοντέλα.
- Με βάση λοιπόν το κριτήριο το απλό μοντέλο με μόνο τη $X_1$ ως επεξηγηματική είναι προτιμότερο παρόλο που έχει κατά πολύ μικρότερο συντελεστή προσδιορισμού. Γιατί;

# Γιατί;

Αυτό οφείλεται στην ακραία παρατήρηση η οποία έχει πολύ μεγάλα κατάλοιπα και επομένως μεγαλύτερα μοντέλα προσπαθούν απλά να βελτιώσουν την προσαρμογή του μοντέλου ως προς την παρατήρηση αυτή.

# Σχέση *Jackknife Cross − validation*

Η εκτίμηση των παραμέτρων αφήνοντας έξω μια παρατήρηση κάθε φορά ξεκάθαρα οδηγεί σε μια *jackknife* μέθοδο εκτίμησης. Στην συγκεκριμένη περίπτωση θα μπορούσε να χρησιμοποιηθούν οι μέσοι των $\hat{a}_{(i)}, \hat{\beta}_{(i)}$ ως εκτιμήσεις των συντελεστών του μοντέλου βασισμένοι στη μέθοδο *jackknife*. Μια τέτοια προσέγγιση θα διόρθωνε το πρόβλημα με την παρατήρηση ακραία που έχει μεγάλη επίδραση στο αποτέλεσμα.