Παλινδρόμηση – Ανάλυση Διασποράς

Πρόβλημα 1

Το αρχείο prob1.dta περιέχει δεδομένα σχετικά με την απόσταση από την ακτή σε km (distance) και τη σκελετική πυκνότητα (σε gr/κ.εκ.) ενός συγκεκριμένου κοραλλιού (density) σε 27 μετρήσεις που έγιναν σε διαφορετικά σημεία του Μεγάλου Κοραλλιογενούς Υφάλου της Βόρειας Αυστραλίας.

1. Να τρέξετε το μονοπαραγοντικό μοντέλο με εξαρτημένη μεταβλητή την density και να αξιολογήσετε τη στατιστική σημαντικότητα.
2. Να δώσετε τις ερμηνείες των εκτιμήσεων των συντελεστών παλινδρόμησης.
3. Δώστε ένα διάστημα πρόβλεψης 95% για τη σκελετική πυκνότητα για μια περιοχή σε απόσταση 18km από την ακτή.
4. Να υπολογίσετε τα κατάλοιπα jackknife, τη μόχλευση (leverage) και την απόσταση Cook για όλες τις παρατηρήσεις. Με βάση αυτά βρείτε (σε επίπεδο σημαντικότητας 5%) ποιες παρατηρήσεις θεωρούνται outliers και ποιες influential.
5. Να πραγματοποιήσετε έλεγχο κανονικότητας των καταλοίπων.

Πρόβλημα 2

Το αρχείο prob2.dta περιέχει δεδομένα σχετικά με το λάθος καταμέτρησης στην απογραφή του 1980 στις ΗΠΑ. Συγκεκριμένα μετά την απογραφή έγιναν κάποιοι δειγματοληπτικοί έλεγχοι σε 66 περιοχές, από τους οποίους προέκυψε ένα εκτιμώμενο ποσοστό «υποκαταμέτρησης» (δηλαδή αν ο πραγματικός πληθυσμός είναι 100 άτομα και η απογραφή καταμετρήσει 95, η υποκαταμέτρηση είναι 5%, ενώ αν η απογραφή καταμετρήσει 103, η υποκαταμέτρηση είναι -3%. Στις ΗΠΑ η θετική υποκαταμέτρηση θεωρείται πρόβλημα για μια περιοχή γιατί επηρεάζει αρνητικά τη χρηματοδότησή της από την ομοσπονδιακή κυβέρνηση). Στα δεδομένα εκτός από το ποσοστό υποκαταμέτρησης, περιέχονται και δεδομένα οικονομικά, κοινωνικά και μορφωτικά. Συγκεκριμένα οι μεταβλητές σημαίνουν τα εξής:

area	Όνομα Περιοχής
perc_min	Ποσοστό μαύρων ή ισπανόφωνων κατοίκων
crimrate	Εγκληματικότητα (Εγκλήματα ανά 1000 κατοίκους)
poverty	Ποσοστό κάτω από το όριο φτώχειας
diffeng	Ποσοστό με δυσκολία στα αγγλικά
hsgrad	Ποσοστό ενηλίκων που δεν έχουν τελειώσει γυμνάσιο
housing	Ποσοστό κατοικιών που είναι διαμερίσματα σε μικρές πολυκατοικίες
undcount	Εκτιμώμενο ποσοστό υποκαταμέτρησης κατά την απογραφή

Σκοπός της μελέτης είναι να εκτιμηθεί κατά πόσο η υποκαταμέτρηση σχετίζεται με κάποιους από τους παραπάνω παράγοντες και ποιους. Σε όλες τις ερωτήσεις η εξαρτημένη μεταβλητή είναι η undcount.

1. Θεωρήστε το πλήρες μοντέλο που περιέχει όλες τις ανεξάρτητες μεταβλητές (εκτός από την area). Στο μοντέλο αυτό ερμηνεύστε τους συντελεστές παλινδρόμησης για όσες μεταβλητές είναι στατιστικά σημαντικές στο επίπεδο 5%.
2. Ξεκινώντας από το πλήρες μοντέλο εκτελέστε διαδοχικά βήματα backward έως ότου καμιά μεταβλητή να μη μπορεί να απομακρυνθεί. Χρησιμοποιήστε p(remove) = 0.05. Γράψτε την εξίσωση παλινδρόμησης για το μοντέλο που καταλήξατε.
3. Έστω το Μοντέλο 2 που περιέχει μόνο τις perc_min, poverty, hsgrad. Απαντήστε στις παρακάτω ερωτήσεις:
    a. Πώς συγκρίνεται το Μοντέλο 2 με το πλήρες μοντέλο από την άποψη της στατιστικής σημαντικότητας; Περιγράψτε το στατιστικό έλεγχο που θα χρησιμοποιήσετε (υποθέσεις και αποτέλεσμα).
    b. Υπολογίστε την τιμή του στατιστικού Cp-Mallows για το Μοντέλο 2 ως προς το πλήρες μοντέλο. Τι συμπεραίνετε;

4. Συγκρίνετε τη στατιστική σημαντικότητα της housing στο μονοπαραγοντικό μοντέλο και στο μοντέλο που περιέχει τις housing και poverty. Πώς εξηγείτε την τυχόν διαφορά; Δικαιολογήστε την εξήγησή σας με κατάλληλη ανάλυση.

Πρόβλημα 3

Το αρχείο prob3.dta περιέχει δεδομένα σχετικά με το ποσοστό κωνοφόρων δέντρων που έπαθαν βλάβες έπειτα από μια επιδημία προσβολής από ένα συγκεκριμένο μήκυτα, σε 64 ορεινές επαρχίες των ΗΠΑ:

region: Περιοχή (βόρειες ή νότιες πολιτείες)
elev: Υψόμετρο (σε m)
damage: Ποσοστό επί τοις 100 δέντρων με βλάβη

Έστω m η μέση τιμή του υψομέτρου όλων των περιοχών και elevc=elev-m η κεντρικοποιημένη τιμή του υψομέτρου. Θεωρήστε το μοντέλο παλινδρόμησης με εξαρτημένη τιμή την damage και ανεξάρτητες τις elevc και region, με κύριες επιδράσεις και αλληλεπιδράσεις, δηλαδή

$$\text{damage} = \beta_0 + \beta_1 \text{ elevc} + \beta_2 \text{ region} + \beta_3 \text{ elevc*region}$$

1. Ποια είναι η εξίσωση παλινδρόμησης για κάθε περιοχή;
2. Ερμηνεύστε τις εκτιμήσεις ελαχίστων τετραγώνων των β_0, β_1, β_2,. β_3,
3. Ποιοι συντελεστές δεν είναι στατιστικά σημαντικοί στο επίπεδο 5%? Με βάση αυτό, ποια είναι η μορφή των εξισώσεων παλινδρόμησης για κάθε κατηγορία; Παραστήστε τις γραφικά (προσεγγιστικά στο γραπτό σας).