Δεύτερη Εργασία Γραμμικών και Γενικευμένων Γραμμικών Μοντέλων, 2026, ΠΜΣ, ΣΑΧΜ, Πανεπιστήμιο Αιγαίου.

Σημείωση: Η εργασία καλό είναι να γίνει σε latex. Μη γράφετε εξισώσεις σε Word. Όπου ζητείται στατιστικός έλεγχος να γράφετε αναλυτικά τη μηδενική υπόθεση, την εναλλακτική υπόθεση, το μοντέλο όπου αναφέρεται ο έλεγχος και το επίπεδο σημαντικότητας.

1. Χρησιμοποιώντας την R, προσομοιώστε δεδομένα από ένα απλό μοντέλο τυχαίων συντελεστών (τυχαίων ευθειών) για 100 άτομα και 14 συνεχείς ημέρες. Το μέσο (πληθυσμιακό) intercept να είναι ίσο με $\beta_0 = 50$ και η (πληθυσμιακή) μέση κλίση να είναι ίση με $\beta_1 = 1$. Οι διακυμάνσεις να είναι: για τα (subject-specific) intercepts $\sigma_0^2 = 4$, για τις (subject-specific) κλίσεις $\sigma_1^2 = 2$, και για το σφάλμα παρατήρησης $\sigma^2 = 9$. Η συν-διακύμανση μεταξύ των intercepts και των κλίσεων να είναι ίση με $\sigma_{01} = -0.05$.
   (α′) Δώστε κατάλληλα γραφήματα για τα δεδομένα που γεννήσατε.
   (β′) Χρησιμοποιώντας τις κατάλληλες εντολές στην R, προσαρμόσετε το κατάλληλο μεικτό μοντέλο και εκτιμήσετε τις παραμέτρους.
   (γ′) Δώστε ένα 95% δ.ε. για το μέσο (πληθυσμιακό) intercept.
   (δ′) Δώστε ένα 95% δ.ε. για τη (πληθυσμιακή) μέση κλίση.
   (ε′) Κάνετε ελέγχο της υπόθεση έλλειψης χρονικής επίδρασης στην μέση απόκριση σε ε.σ. 5%.
   (ϛ′) Δώστε ένα 95% δ.ε. για τη μέση (πληθυσμιακή) τιμή της απόκρισης την 6η ημέρα.

2. (Συνέχεια της προηγούμενης) Προσομοιώστε δεδομένα από ένα μοντέλο τυχαίων συντελεστών για 200 άτομα και δέκα ημέρες. Τα 100 άτομα να είναι στο placebo group και τα υπόλοιπα 100 άτομα στο drug group. Το μέσο intercept για το placebo group να είναι 50 και η μέση κλίση για το placebo group να είναι 1. Το μέσο intercept για το drug group να είναι 50 και η μέση κλίση για το drug group να είναι 1.50. Οι διακυμάνσεις να είναι: για τα intercept ίση με 4, για τις κλίσεις ίση με 2 , για το σφάλμα παρατήρησης ίση με 9 και η συνδιακύμανση μεταξύ intercepts και κλίσεων να είναι -0.05.
   (α′) Δώστε κατάλληλα γραφήματα για τα δεδομένα που γεννήσατε.
   (β′) Χρησιμοποιώντας τις κατάλληλες εντολές στην R, προσαρμόσετε το μοντέλο και εκτιμήσετε τις παραμέτρους.
   (γ′) Προσαρμόσετε ένα μοντέλο που να μη λαμβάνει καθόλου υπόψη το φάρμακο (θεραπεία).
   (δ′) Ελέγξτε την υπόθεση που αναφέρει πως το φάρμακο δεν επηρεάζει την μέση απόκριση.
   (ε′) Ελέγξτε την υπόθεση που αναφέρει πως το φάρμακο δεν επηρεάζει την χρονική πορεία της μέσης απόκρισης. Ο έλεγχος να γίνει με χρήση του Likelihood ratio test statistic.
   (ϛ′) Δώστε 95% δ.ε. για την διαφορά των μέσων κλίσεων στις 2 ομάδες.
   (ζ′) Δώστε 95% δ.ε. για την διαφορά της μέσης απόκρισης μεταξύ των 2 ομάδων κατά την 9η ημέρα.
   (η′) Πολλοί ερευνητές ισχυρίζονται πως είναι λογικό στο μοντέλο να μην υπάρχει διαφορά στο μέσο intercept μεταξύ των 2 ομάδων. Ποιά είναι η γνώμη σας; Κάνετε έναν έλεγχο της υπόθεσης κοινού intercept βάσει των δεδομένων σας.

3. Θεωρήστε πως η μέση τιμή ημερήσιων θανάτων σε μια χώρα από μία πανδημία ακολουθεί το μοντέλο:

$$\mu_t = \frac{9000\sqrt{2\pi}}{45}\phi\left(\frac{t-100}{45}\right),\ t = 1, 2, \ldots$$

όπου t αναφέρεται στην ημέρα, και η συνάρτηση $\phi()$ είναι η συνάρτηση πυκνότητας πιθανότητας μιας τυπικής κανονικής κατανομής. Για τους παρατηρούμενους ημερήσιους αριθμούς θανάτων, Y_t , υποθέσετε πως είναι ανεξάρτητοι μεταξύ τους και ακολουθούν την $Poisson(\mu_t)$ κατανομή.

(α′) Δείξτε πως το παραπάνω μοντέλο είναι ένα γενικευμένο γραμμικό μοντέλο. Τι ιδιαίτερο έχει;

(β′) Σε ποια ημέρα αναμένουμε τους περισσότερους θανάτους; Ποια η αναμενόμενη τιμή των θανάτων σε αυτή την μέρα;

(γ′) Χρησιμοποιώντας την R, προσομοιώσετε τιμές ημερησίων θανάτων, για $t = 1, 2, \ldots, 120$.

(δ′) Προσαρμόστε το κατάλληλο γενικευμένο γραμμικό μοντέλο στα δεδομένα. Εκτιμήσετε την αναμενόμενη τιμή ημερησίων θανάτων για την 121η ημέρα και δώστε ένα 90% διάστημα εμπιστοσύνης για αυτήν. Σχολιάστε!

(ε′) Δώστε κατάλληλο γράφημα για τα δεδομένα που γεννήσατε όπου θα φαίνεται η πραγματική πορεία της πανδημίας, η εκτιμώμενη πορεία και οι παρατηρηθείσες τιμές ημερήσιων θανάτων.

4. Θεωρήστε πως η τιμή ενός συνεχούς βιοδείκτη, X, είναι ενδεικτική της ύπαρξης μιας ασθένειας (μεγαλύτερη τιμή σημαίνει 'περισσότερη' ένδειξη ασθένειας). Έστω Y η δείκτρια ασθένειας (=1 αν το άτομο βρέθηκε όντως ασθενής). Δίνεται πως $P(Y = 1) = \pi$ και πως $X|Y = 0 \sim F_0$ ενώ $X|Y = 1 \sim F_1$, δηλαδή η κατανομή του βιοδείκτη για τους υγιείς είναι F_0 και για τους ασθενείς η κατανομή είναι F_1. Συμβολίζουμε με f_0 και f_1 τις σ.π.π. των υγιών και των ασθενών αντίστοιχα.

(α′) Βρείτε την $P(Y = 1|X = x)$ σαν συνάρτηση του λόγου $r(x) = \frac{f_1(x)}{f_0(x)}$.

(β′) Βρείτε το $logit(P(Y = 1|X = x))$. Δείξτε πως όταν θεωρήσουμε το Y ως απόκριση έχουμε να κάνουμε με λογιστική παλινδρόμηση.

(γ′) Βρείτε την λογιστική παλινδρόμηση όταν $\pi = 0.30$, $F_0 = N(10, 1)$ και $F_1 = N(12, 4)$.

(δ′) Δώστε το γράφημα της συνάρτησης της πιθανότητας ασθένειας με τις παραμέτρους που βρήκατε.

(ε′) Χρησιμοποιώντας την R, προσομοιώσετε αρχικά 400 ανεξάρτητες τιμές $y_i \sim Bernoulli(0.30)$. Μετά για $y_i = 0$ προσομοιώστε τιμή $x_i \sim N(10, 1)$ ενώ για $y_i = 1$ προσομοιώστε τιμή $x_i \sim N(12, 4)$.

(ϛ′) Χρησιμοποιώντας τα δεδομένα που γεννήσατε, προσαρμόστε την κατάλληλη λογιστική παλινδρόμηση και εκτιμήσετε, με χρήση R, τις παραμέτρους. Ελέγξετε τη στατιστική σημαντικότητα της παλινδρόμησης σε ε.σ. 5%.

(ζ′) Δώστε το γράφημα της εκτίμησης της συνάρτησης της πιθανότητας ασθένειας με τις εκτιμήσεις που βρήκατε. Το γράφημα αυτό να δοθεί μαζί με το γράφημα στο (δ) σε κοινό plot.

Σημείωση: Στις προσομοιώσεις σας να ξεκινάτε με seed= ημερομηνία γέννησής σας και προσθέτετε +1 για την επόμενη.

BONUS : Στα Ερωτήματα 1 και 2 βρείτε τις εκτιμήσεις των παραμέτρων μεγιστοποιώντας απευθείας τις πιθανοφάνειες (αφού πρώτα τις ορίσετε στη R, σύμφωνα με τις σημειώσεις). Σχολιάστε.