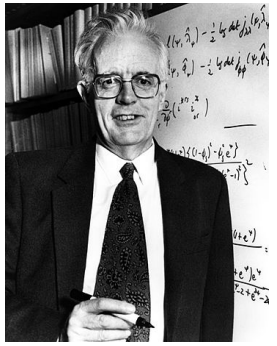
Στατιστική στην Πληροφορική

13. Λογιστική Παλινδρόμηση

Περιεχόμενα

- Συμπερασματολογία
  - Διαστήματα εμπιστοσύνης
  - Έλεγχος σημαντικότητας
- Πρόβλεψη
- Παραδείγματα

Λογιστική παλινδρόμηση



- Logistic regression (D. Cox, 1958)

- Συμπερασματολογία για τη σχέση *ποσοτικής επεξηγηματικής* μεταβλητής με *κατηγορική μεταβλητή απόκρισης*
  - Πχ, τι σχέση έχουν οι ώρες μελέτης ενός μαθήματος με την επιτυχία στο μάθημα αυτό;
  - Πχ, σχέση έκθεσης στον ήλιο με εμφάνιση μελανώματος

- Χρησιμοποιείται ως *μέθοδος πρόβλεψης* στη Μηχανική Μάθηση:
  - Πχ, αναγνώριση προτύπων

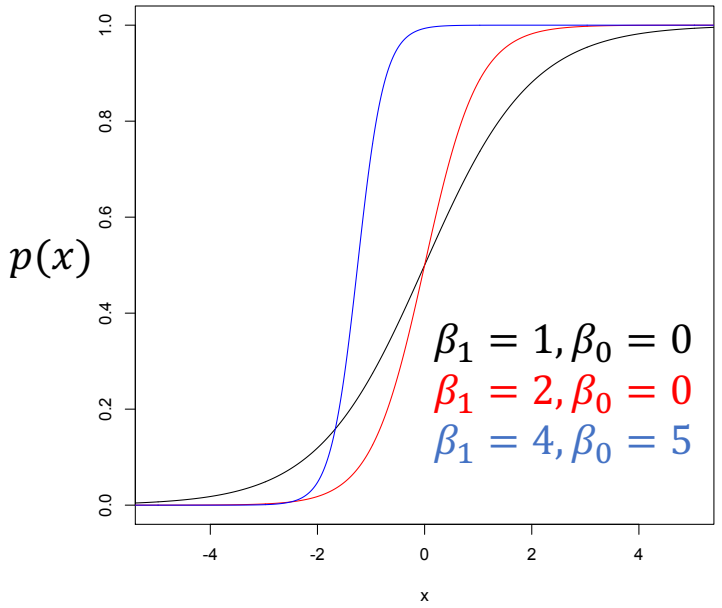
Λογιστική παλινδρόμηση

- X: *ποσοτική επεξηγηματική* μεταβλητή
  - Εάν δίτιμη κατηγορική, μετατρέπεται σε ποσοτική με τιμές {0,1}
- Y: *δίτιμη* κατηγορική μεταβλητή απόκρισης, {0,1} από υπόθεση
  - Εάν λαμβάνει πάνω από 2 τιμές χρησιμοποιείται πολυωνυμική λογιστική παλινδρόμηση (multinomial logistic regression)
- *Παράμετρος πληθυσμού*: $p(x) =$ ποσοστό '1' στον υποπληθυσμό όπου $X = x$
- Λόγος πιθανοτήτων (*odds*) '1' προς '0': $\frac{p(x)}{1-p(x)}$
- Λογάριθμος odds: $\log\left(\frac{p(x)}{1-p(x)}\right)$
- Υπόθεση για πληθυσμό: $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_1 x + \beta_0$ για κάθε x

  ή ισοδύναμα, $p(x)/(1-p(x)) = e^{\beta_1 x+\beta_0}$

  ή ισοδύναμα, $p(x) = \frac{1}{1+e^{-(\beta_1 x+\beta_0)}}$

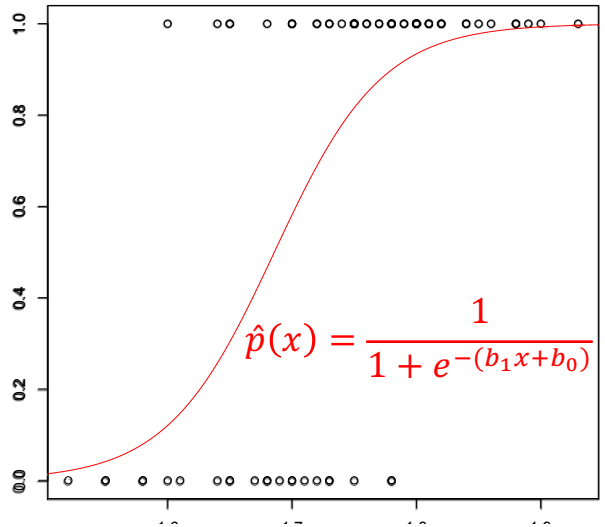
- *Odds ratio*: $e^{\beta_1} = \frac{\frac{p(x+1)}{1-p(x+1)}}{\frac{p(x)}{1-p(x)}}$



Συμπερασματολογία

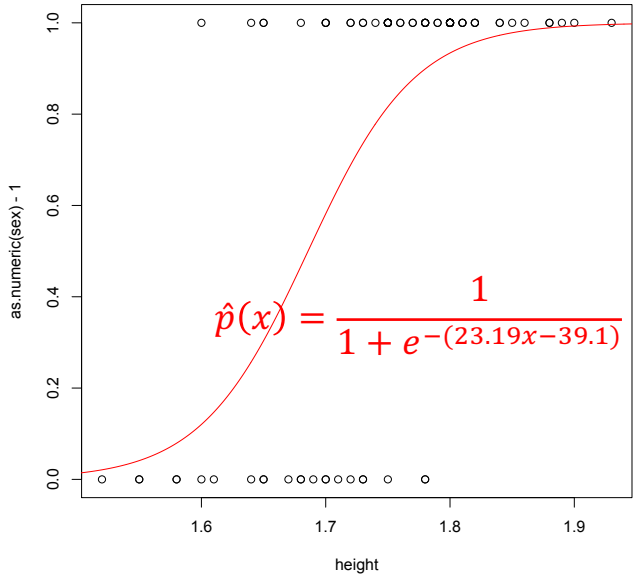
- Δεδομένα: για οποιεσδήποτε n γνωστές τιμές $x_1, \dots, x_n$ της X
  - Από κάθε υποπληθυσμό με $X = x_i$, λαμβάνουμε ένα τυχαίο δείγμα y_i μεγέθους 1

- Εκτιμητές
  - $\hat{p}(x)$: εκτιμητής ποσοστού υποπληθυσμού x
  - b_1: εκτιμητής της κλίσης β_1
  - b_0: εκτιμητής της σταθεράς β_0
  - Υπολογισμός από λογισμικό



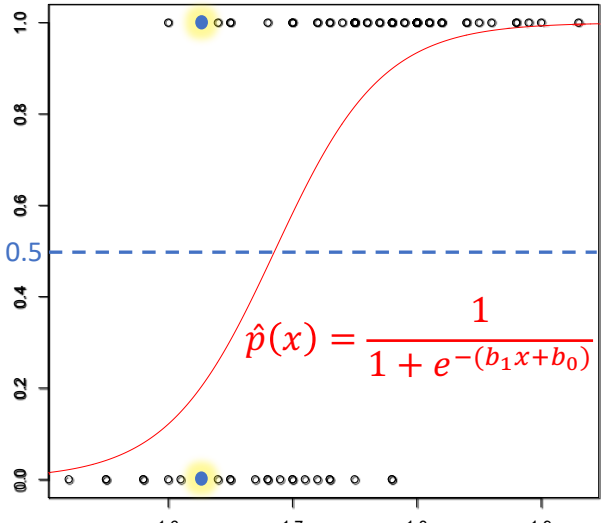
Παράδειγμα

- Ποια σχέση έχει το ύψος (επεξηγηματική) με το φύλο (απόκριση) στους φοιτητές πληροφορικής;
- Απλό τυχαίο δείγμα (από ερωτηματολόγιο): $n = 80$
- $b_1 = 23.19, b_0 = -39.1$



Πρόβλεψη (Prediction)

- Δεδομένα: για οποιεσδήποτε n γνωστές τιμές $x_1, \dots, x_n$ της X
  - Από κάθε υποπληθυσμό με $X = x_i$, λαμβάνουμε ένα τυχαίο δείγμα y_i μεγέθους 1
- Νεό δεδομένο: (x, y)
  - x παρατηρήσιμο, y άγνωστο: *πρόβλεψη* y?
- Εκτιμητές
  - $\hat{p}(x)$: εκτιμητής ποσοστού υποπληθυσμού x
- Πρόβλεψη: $\hat{y} = \begin{cases} 1, & \hat{p}(x) > 1/2 \\ 0, & \hat{p}(x) < 1/2 \end{cases}$

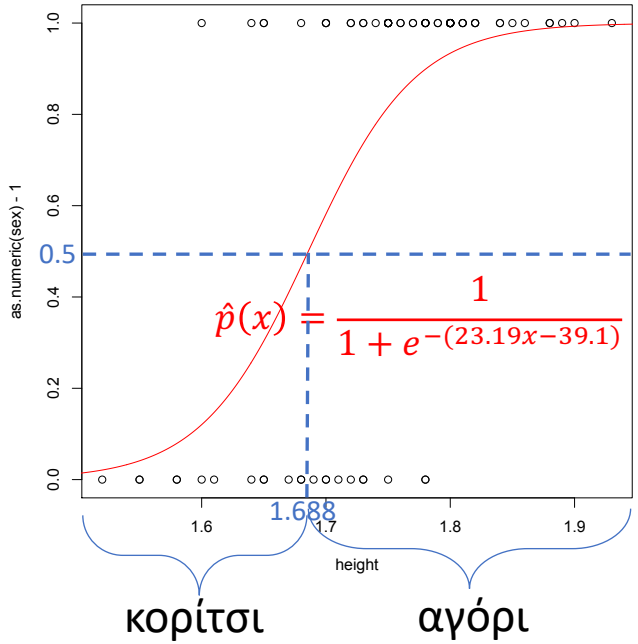


Παράδειγμα

- Ποια σχέση έχει το ύψος (επεξηγηματική) με το φύλο (απόκριση) στους φοιτητές πληροφορικής;
- Απλό τυχαίο δείγμα (από ερωτηματολόγιο): $n = 80$

- $b_1 = 23.19, b_0 = -39.1$

- Πρόβλεψη: $\hat{y} = \begin{cases} \text{αγορι}, & x > 1.688 \\ \text{κοριτσι}, & x < 1.688 \end{cases}$



Εκτιμητές

- Απλός υπολογισμός των εκτιμητών b_0, b_1 για δίτιμη *επεξηγηματική* μεταβλητή $X \in \{0,1\}$
  - Πχ, σχέση φύλου και επιτυχίας στις Πιθανότητες στους φοιτητές πληροφορικής που θα παίρνουν το μάθημα της Στατιστικής στην Πληροφορική;

- Έστω $p(A), p(K)$ ποσοστό επιτυχίας σε αγόρια και κορίτσια αντίστοιχα

- Υπόθεση για πληθυσμό:
  - $\log\left(\frac{p(A)}{1-p(A)}\right) = \beta_1 \cdot 1 + \beta_0$
  - $\log\left(\frac{p(K)}{1-p(K)}\right) = \beta_1 \cdot 0 + \beta_0$

Εκτιμητές

- Απλός υπολογισμός των εκτιμητών b_0, b_1 για δίτιμη *επεξηγηματική* μεταβλητή $X \in \{0,1\}$
  - Πχ, σχέση φύλου και επιτυχίας στις Πιθανότητες στους φοιτητές πληροφορικής που θα παίρνουν το μάθημα της Στατιστικής στην Πληροφορική;
- Δεδομένα: ερωτηματολόγιο
  - Λόγος επιτυχίας/αποτυχίας στα αγόρια: $\frac{46}{11} = 4.18$
  - Λόγος επιτυχίας/αποτυχίας στα κορίτσια: $\frac{16}{8} = 2$
- Γραμμική παλινδρόμηση:
  - $\log(4.18) = b_1 \cdot 1 + b_0$
  - $\log(2) = b_1 \cdot 0 + b_0$
- Άρα, $b_1 = \log(2.09) = 0.737, b_0 = \log(2) = 0.693$
- Εκτιμητής odds ratio = δειγματικό odds ratio = $e^{b_1} = 2.09 = \frac{\frac{46}{11}}{\frac{16}{8}}$

	Κορίτσι	Αγόρι
FALSE	8	11
TRUE	16	46

Συμπερασματολογία: διαστήματα εμπιστοσύνης

- $C\%$ διάστημα εμπιστοσύνης για *κλίση* β_1: $b_1 \pm z_* SE_{b_1}$
- $C\%$ διάστημα εμπιστοσύνης για *odds ratio* e^{β_1}: $e^{b_1 \pm z_* SE_{b_1}}$
  - z_* δίνεται από την τυπική κανονική κατανομή
  - SE_{b_1}: δειγματική τυπική απόκλιση εκτιμητή
    - Υπολογίζεται από λογισμικό

- Πχ, 95% διάστημα εμπιστοσύνης για odds ratio επιτυχίας/αποτυχίας σε αγόρια προς κορίτσια:
  1. Από δεδομένα: $b_1 = 0.737, SE_{b_1} = 0.548$
  2. $z_* = 1.96$
  3. $e^{z_* SE_{b_1}} = 2.927$
  4. Διάστημα = $\left[\frac{2.09}{2.927}, 2.09 \times 2.927\right] = [0.7, 6.13]$

Συμπερασματολογία: έλεγχος σημαντικότητας

- Έλεγχος για την ύπαρξη σχέσης μεταξύ των μεταβλητών:

  $H_0: \beta_1 = 0$

  $H_a: \beta_1 \neq 0$

- Στατιστικό ελέγχου (*Wald statistic*) $z = \frac{b_1}{SE_{b_1}}$
- *p value* δίνεται από τυπική κανονική κατανομή
- Πχ, έχει σχέση το ύψος με το φύλο;
  1. Από δεδομένα: $b_1 = 23.18, SE_{b_1} = 5.41$
  2. $z = \frac{23.18}{5.41} = 4.289$
  3. p value $= 1.8 \times 10^{-5} \Rightarrow$ απορρίπτεται η $\beta_1 = 0$
- Πχ, έχει σχέση το φύλο με την επιτυχία στις Πιθανότητες;
  - p value=0.178. (Προσοχή! Δεν είναι ισοδύναμος έλεγχος με δίπλευρο z για σύγκριση ποσοστών μεταξύ δύο πληθυσμών ή χ^2, όπου δίνουν 0.173)