

## Στατιστικές Μέθοδοι στην Επιδημιολογία

### Lab 5

**Σκοπός:** Εφαρμογή σταθμισμένων μεθόδων για την εκτίμηση επιπολασμού και τη μοντελοποίηση παραγόντων κινδύνου, με έμφαση στην ορθή ερμηνεία υπό σύνθετο δειγματοληπτικό σχεδιασμό.

#### 1. Προετοιμασία Δεδομένων & Έλεγχος Ελλειπουσών Τιμών

1. Δηλώστε εκ νέου το δειγματοληπτικό σχέδιο χρησιμοποιώντας τα τελικά βάρη που δημιουργήσατε στο Lab 4:

```
. svyset blockid [pweight = weight_final], strata(strata)
singleunit(centered)
```

2. Ενώστε το κύριο dataset με το αρχείο εξετάσεων data\_exams.dta χρησιμοποιώντας το μοναδικό αναγνωριστικό emenoid.

```
. merge 1:1 emenoid using "data_exams.dta"
```

3. Ποιο ποσοστό των συμμετεχόντων δεν έχει δεδομένα εξετάσεων; Τι μπορεί να υποδηλώνει αυτό;

```
. tab _merge
```

4. Εξετάστε τα μοτίβα ελλειπουσών τιμών για τις βασικές μεταβλητές (π.χ. ηλικία, φύλο, BMI, υπέρταση, υπερχοληστερολαιμία, φυσική δραστηριότητα).

```
. misstable summarize
. misstable patterns age gender bmi_doct PA urban diab_prev
hypertension_pr hyperchol200 hyperldl130
```

**\* (Optional)**

```
ssc install mdesc, replace
. mdesc age gender bmi_doct PA urban diab_prev hypertension_pr
hyperchol200 hyperldl130
```

5. Συζητήστε τον πιθανό μηχανισμό απουσίας δεδομένων (MCAR, MAR, MNAR). Πώς μπορεί να επηρεάσει τη μεροληψία των εκτιμήσεων;

## 2. Ανάλυση συμμετοχής στις εξετάσεις

6. Δημιουργήστε μεταβλητή συμμετοχής στις εξετάσεις (participated) και εφαρμόστε λογιστικό μοντέλο για να εντοπίσετε παράγοντες που σχετίζονται με τη συμμετοχή. Ποιοι συμμετείχαν περισσότερο; Υπάρχει πιθανό selection bias;

```
. gen participated = (_merge == 3)
. logistic participated i.gender i.age_gr i.urban i.area
```

7. Ερμηνεύστε τα αποτελέσματα:

- Ποια ηλικιακή ομάδα είχε τη μεγαλύτερη πιθανότητα συμμετοχής;
- Υπήρξαν διαφορές ανά φύλο ή περιοχή;
- Πώς μεταβάλλεται η πιθανότητα συμμετοχής ανάλογα με τον βαθμό αστικότητας;

8. Ποιες μέθοδοι μπορούν να χρησιμοποιηθούν για να περιορίσουν τη μεροληψία λόγω διαφορεικής συμμετοχής;

9. Στο υπόλοιπο του Lab, εφαρμόστε **complete-case analysis** (κρατήστε μόνο όσους έχουν δεδομένα εξετάσεων):

```
. drop if _merge == 1
. drop _merge
```

## 3. Σταθμισμένη Περιγραφική Ανάλυση

10. Υπολογίστε σταθμισμένα περιγραφικά για ηλικία, φύλο, περιοχή, αστικότητα, φυσική δραστηριότητα και δείκτη μάζας σώματος.

```
. svy: tab variable, per obs

. svy: mean variable
. estat sd
```

**11.** Εκτιμήστε τον σταθμισμένο επιπολασμό (prevalence) και το 95% διάστημα εμπιστοσύνης για:

- Υπέρταση (hypertension\_pr)
- Σακχαρώδη διαβήτη (diab\_pr)
- Υπερχοληστερολαιμία (hyperchol200)

```
. svy: tab hypertension_pr, per obs ci  
. svy: tab diab_prev, per obs ci  
. svy: tab hyperchol200, per obs ci
```

**12.** Συγκρίνετε τους τρεις δείκτες. Ποιος δείκτης έχει τη μεγαλύτερη συχνότητα και τι μπορεί να υποδηλώνει αυτό για το δείγμα;

#### **4. Σταθμισμένα Μοντέλα Λογιστικής Παλινδρόμησης**

##### **4.1 Υπέρταση και BMI**

**13.** Εξετάστε αν η **υπέρταση** σχετίζεται με το BMI.

```
. svy: logit hypertension_pr i.bmi_doc_cat, or
```

**14.** Εκτελέστε επιμέρους (μονοπαραγοντικές) αναλύσεις για να εντοπίσετε ποιοι παράγοντες σχετίζονται σημαντικά με την υπέρταση:

```
. svy: logit hypertension_pr i.gender, or  
. svy: logit hypertension_pr age, or  
. svy: logit hypertension_pr i.urban, or  
. svy: logit hypertension_pr alcohol, or
```

**15.** Δημιουργήστε το πολυπαραγοντικό μοντέλο, προσθέτοντας διαδοχικά τις μεταβλητές: ηλικία, φύλο, αστικότητα και BMI.

**16.** Ποια μεταβλητή φαίνεται να δρα ως συγχυτικός παράγοντας; Επαναλάβετε το μοντέλο εισάγοντας τις μεταβλητές μία προς μία:

#### 4.2 Υψηλή LDL

17. Εκτιμήστε τον επιπολασμό ατόμων με LDL >130 mg/dL (hyperldl130) και συγκρίνετέ τον ανά φύλο.

```
. svy: tab hyperldl130, per obs ci  
. svy: tab hyperldl130 gender, per obs col
```

18. Εφαρμόστε σταθμισμένο λογιστικό μοντέλο με ανεξάρτητες μεταβλητές ηλικία, φύλο και BMI. Ποιοι παράγοντες σχετίζονται στατιστικά με υψηλή LDL; Πώς ερμηνεύονται οι συντελεστές;

```
. svy: logistic hyperldl130 i.gender c.age c.bmi
```

19. Εξετάστε αν υπάρχει αλληλεπίδραση μεταξύ ηλικίας και φύλου:

```
. svy: logistic hyperldl130 i.gender##c.age i.bmi_doc_cat
```

20. Υπολογίστε και ερμηνεύστε τα αποτελέσματα για συγκεκριμένες ηλικίες (π.χ. 20, 35, 50, 70, 85 έτη) χρησιμοποιώντας lincom:

```
. lincom _b[1.gender] + 20*_b[1.gender#c.age], eform  
. lincom _b[1.gender] + 35*_b[1.gender#c.age], eform  
. lincom _b[1.gender] + 50*_b[1.gender#c.age], eform  
. lincom _b[1.gender] + 70*_b[1.gender#c.age], eform  
. lincom _b[1.gender] + 85*_b[1.gender#c.age], eform
```

21. Δημιουργήστε διάγραμμα με προβλεπόμενες πιθανότητες ανά φύλο και ηλικία:

```
. margins gender, at(age=(18(10)100))  
. marginsplot
```

22. Εξετάστε αν η σχέση ηλικίας–LDL είναι γραμμική ή μη γραμμική. Δημιουργήστε διαφορετικές λειτουργικές μορφές της ηλικίας.

```
. gen age2 = age^2  
. gen age3 = age^3  
. gen age12 = age^(1/2)  
. rc_spline age
```

23. Εφαρμόστε εναλλακτικά μοντέλα (γραμμικό, τετραγωνικό, κυβικό, τετραγωνική ρίζα, spline) και συγκρίνετε την καταλληλότητα:

```
. svy: logit hyperldl age  
. svy: logit hyperldl age age2  
. svy: logit hyperldl age age2 age3  
. svy: logit hyperldl age age12  
. svy: logit hyperldl _S*
```

24. Οπτικοποιήστε τις προβλεπόμενες τιμές (logits) για κάθε μορφή:

```
. twoway scatter pr_linear age || scatter pr_cubic age || scatter  
pr_spline age || scatter pr_sqrt age
```

25. Ποια μορφή περιγράφει καλύτερα τη σχέση; Είναι η σχέση ηλικίας–LDL απολύτως γραμμική;

26. Δημιουργήστε το τελικό μοντέλο με τετραγωνικό όρο της ηλικίας και αλληλεπιδράσεις φύλου–ηλικίας:

```
. svy: logit hyperldl c.age##i.gender c.age#c.age##i.gender  
i.bmi_doc_cat  
. margins gender, at(age = (18(10)100)) atmeans  
. marginsplot
```

#### 4.3 Αλληλεπιδράσεις Συνεχών Μεταβλητών

27. Εκτιμήστε την επίδραση της ηλικίας και του BMI (ως συνεχών μεταβλητών) στην υπερχοληστερολαιμία (hyperchol200), συμπεριλαμβάνοντας την αλληλεπίδρασή τους.

```
. svy: logit hyperchol200 i.gender c.bmi_doct##c.age i.walk210
```

28. Υπολογίστε αναλογίες πιθανοτήτων (OR) για συγκεκριμένες τιμές BMI (π.χ. 20, 25, 30, 35, 40) για αύξηση 10 ετών στην ηλικία:

```
. lincom 10*_b[c.age] + 10*20*_b[c.age#c.bmi_doct], eform  
. lincom 10*_b[c.age] + 10*25*_b[c.age#c.bmi_doct], eform
```

```
. lincom 10*_b[c.age] + 10*30*_b[c.age#c.bmi_doct], eform  
. lincom 10*_b[c.age] + 10*35*_b[c.age#c.bmi_doct], eform  
. lincom 10*_b[c.age] + 10*40*_b[c.age#c.bmi_doct], eform
```

**29.** Δημιουργήστε πλέγμα προβλεπόμενων πιθανοτήτων (margins) και απεικονίστε

```
. margins, at(bmi_doct=(10(3)40) age=(20(20)100))  
saving(predictions, replace)  
. use predictions, clear  
. rename _at1 bmi_doct  
. rename _at2 age  
. rename _margin pr_highchol  
. twoway contour pr_highchol bmi_doct age, title("Predicted  
Pr(chol>200) over BMI and age")
```

## **5. Τελικό Μοντέλο & Ερμηνεία**

**30.** Ποια μεταβλητά μοντέλα αποδίδουν καλύτερα; Ποιες αλληλεπιδράσεις είναι στατιστικά σημαντικές;

**31.** Συγκρίνετε τα αποτελέσματά σας με τη διαθέσιμη βιβλιογραφία ή τα επίσημα ευρήματα της μελέτης EMENO. Είναι τα επίπεδα και οι παράγοντες κινδύνου συμβατά με αυτά άλλων ευρωπαϊκών χωρών;

### **Ερώτηση Αναστοχασμού:**

Πώς συνδέεται η χρήση σταθμισμένων μοντέλων (π.χ. svy: logistic) με την εγκυρότητα των επιδημιολογικών συμπερασμάτων;

Ποια μορφή μεροληψίας θα μπορούσε να προκύψει αν αγνοούσαμε τα βάρη ή τη στρωματοποίηση;