ΒΙΟΣΤΑΤΙΣΤΙΚΗ

Εξάρτηση ποσοτικών χαρακτηριστικών: η έννοια της εξάρτησης περισσότερων από δύο μεταβλητών

ΜΙ Κάσδαγλη, PhD

Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής,
Ιατρική Σχολή Πανεπιστημίου Αθηνών

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η Y (εξαρτημένη) από την X (ανεξάρτητη), με τη σχέση:

$$Y = a + b^* X$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3313.770	54.295		61.033	.000	3206.892	3420.648
	wtgain	19.630	4.526	.251	4.337	.000	10.721	28.540

a. Dependent Variable: bweight

$$Y = a + b^* X \Rightarrow \text{bweight} = 3313.7 + 19.6^* \text{wtgain}$$

Ερμηνεία συντελεστή εξάρτησης b (=19.6): Για ένα κιλό αύξηση του βάρους της μητέρας (wtgain) αναμένεται 19,6 γραμμάρια αύξηση στο βάρος γέννησης του νεογνού κατά μέσο όρο (με 95% πιθανότητα η αύξηση να κυμαίνεται από 10.721 έως 28.540 γραμμάρια)

Εξαρτάται στατιστικά σημαντικά σε επίπεδο σημαντικότητας 5% το βάρος γέννησης από το βάρος που πήρε η μητέρα κατά την εγκυμοσύνη; **ΝΑΙ** (p-value <0.05 & 95% ΔΕ δεν περιλαμβάνει το 0)

Με βάση το μοντέλο μπορούμε να βρούμε την αναμενόμενη τιμή της Y για κάθε τιμή της X, αντικαθιστώντας τις τιμές στον τύπο. Για παράδειγμα, μια γυναίκα που πήρε κατά την εγκυμοσύνη **10** κιλά, αναμένεται να γεννήσει νεογνό βάρους bweight = 3313.7 + 19.6* **10** = 3509,7 γραμμαρίων κατά μέσο όρο.

ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ ΠΟΣΟΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

- Στην έναρξη και κατά τη διάρκεια βιο-ιατρικών μελετών συλλέγονται ποικίλα δεδομένα σχετικά με δημογραφικά χαρακτηριστικά (ηλικία, φύλο), ιατρικό ιστορικό κ.α

- Μερικοί παράγοντες μπορεί να σχετίζονται με το δείκτη υγείας υπό μελέτη.

Συγχυτικοί παράγοντες - παράδειγμα

- Συγχυτική δράση μεταξύ έκθεσης, δείκτη και τρίτου (αιτιολογικού) παράγοντα.

Η αναλογία καπνιστών στις συγκρινόμενες ομάδες (αλκοόλ ναι/όχι) μπορεί να διαταράξει τη σχέση
κατανάλωση αλκοόλ→ Ca πνεύμονα

Περισσότεροι καπνιστές καταναλώνουν αλκοόλ. Αν δε λάβω υπόψη το κάπνισμα, θα βρω σχέση κατανάλωσης αλκοόλ και Ca πνεύμονα η οποία όμως θα είναι πλασματική. Στην πραγματικότητα, αυτοί που καταναλώνουν αλκοόλ έχουν μεγαλύτερο κίνδυνο εμφάνισης Ca πνεύμονα γιατί ΟΙ ΠΕΡΙΣΣΟΤΕΡΟΙ ΕΙΝΑΙ ΚΑΠΝΙΣΤΕΣ (αφού το εξετάσουμε πρώτα).

Πολλαπλή γραμμική εξάρτηση/παλινδρόμηση (Multiple linear regression)

- Ιδανικά, θα θέλαμε
  - να αξιολογήσουμε ταυτόχρονα την επίδραση πολλών παραγόντων και το βαθμό στον οποίο εξηγούν τη μεταβλητότητα από άτομο σε άτομο
  - να αξιολογήσουμε την επίδραση κάθε παράγοντα λαμβάνοντας υπόψη και τους άλλους ώστε να αποκλείσουμε σχέσεις που οφείλονται σε συγχυτικούς παράγοντες

→ **Πολλαπλή γραμμική εξάρτηση**

Πολλαπλή γραμμική εξάρτηση

- **Αντικείμενο**

  Διερεύνηση του τρόπου μεταβολής μια μεταβλητής (εξαρτημένης) συναρτήσει των μεταβολών άλλων μεταβλητών (ανεξάρτητων)

  Διερευνάται η **γραμμική** σχέση μιας εξαρτημένης μεταβλητής με περισσότερες από μία ανεξάρτητες μεταβλητές

- **Προϋπόθεση**

  Η εξαρτημένη μεταβλητή (Υ) να κατανέμεται κανονικά για κάθε συνδυασμό τιμών των ανεξάρτητων μεταβλητών (X_i)

Μοντέλο πολλαπλής γραμμικής εξάρτησης

- Έστω $X_1, X_2, \ldots, X_p$ αντιπροσωπεύουν p ανεξάρτητες μεταβλητές , τότε

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \quad \ldots + b_p X_{pi}$$

$\hat{Y}_i$: Προβλεπόμενη τιμή

$b_1, b_2, \ldots, b_p$: Συντελεστές μερικής εξάρτησης

Ερμηνεία μερικών συντελεστών εξάρτησης

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \quad \ldots + b_p X_{pi}$$

Ο συντελεστής b_i εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής (Υ) κατά μέσο όρο

όταν

η αντίστοιχη ανεξάρτητη X_i μεταβληθεί κατά μία μονάδα

και

όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές ή ανεξάρτητα από τις άλλες μεταβλητές

- Σε μελέτη για τη διερεύνηση της επίδρασης του μολύβδου στην σωματομετρική ανάπτυξη των παιδιών, μελετήθηκαν παιδιά σχολικής ηλικίας από τρεις περιοχές:
  - Λαύριο, Ελευσίνα και Λουτράκι

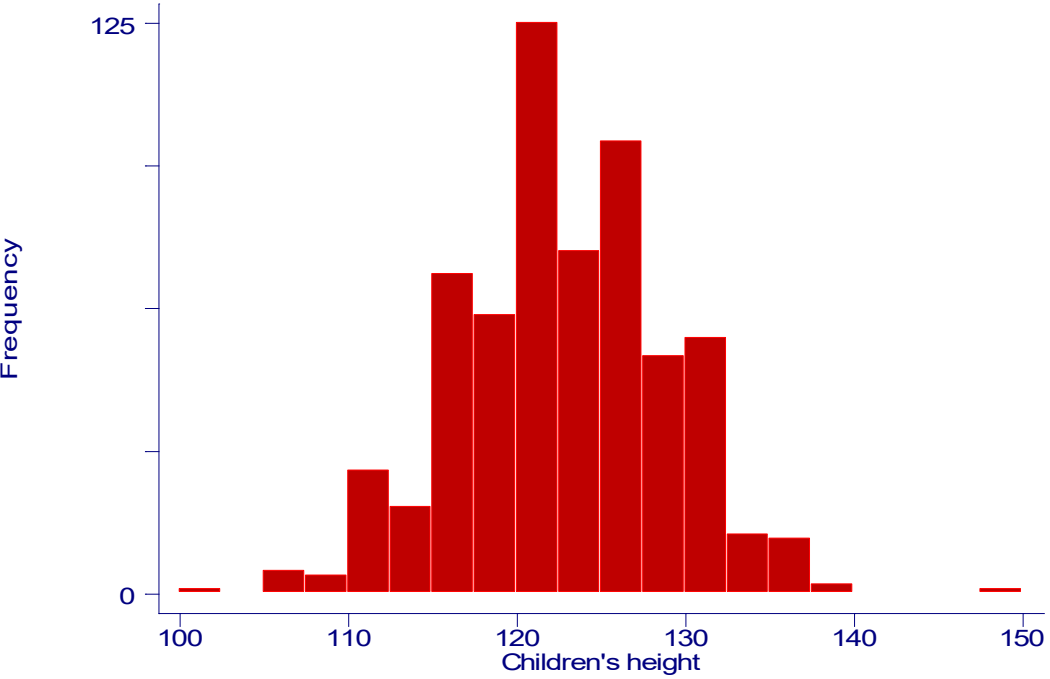
- Το συνολικό δείγμα αποτελείται από 522 παιδιά, 274 αγόρια και 248 κορίτσια ηλικίας 6-9 χρονών. Μέρος των δεδομένων παρουσιάζεται στον πίνακα που ακολουθεί

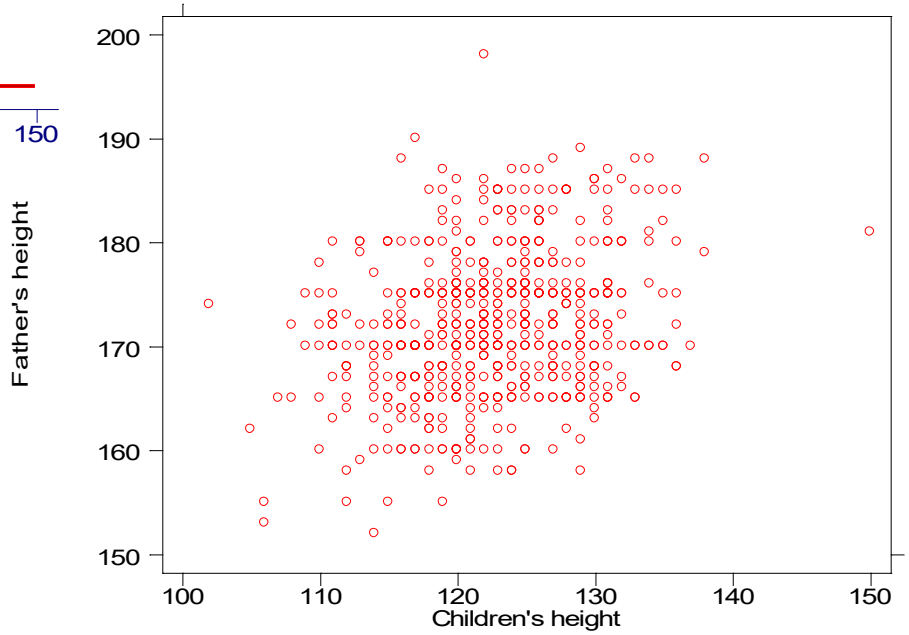
Πηγή: Kafourou et al, 1997.

Κωδικός	Πόλη	Ηλικία (έτη)	Ανάστημα πατέρα (cm)	Μόλυβδος (μg/mL)	Ανάστημα παιδιού (cm)
353	2	8	172	23.42	116
419	2	.	165	51.17	107
19	1	8	152	.	114
26	1	7	177	5.94	122
506	2	7	155	20.21	119
683	3	8	170	4.16	117
612	3	7	164	9.78	112
97	1	8	164	.	121
504	2	7	172	17.29	113
469	2	9	170	26.98	124
498	2	7	160	13.24	110
565	2	8	168	22.94	123
140	1	8	162	2.86	115
374	2	6	155	26.59	112
673	3	6	172	5.69	119
644	3	8	167	11.87	123
507	2	8	177	10.19	125
711	3	7	165	4.15	124

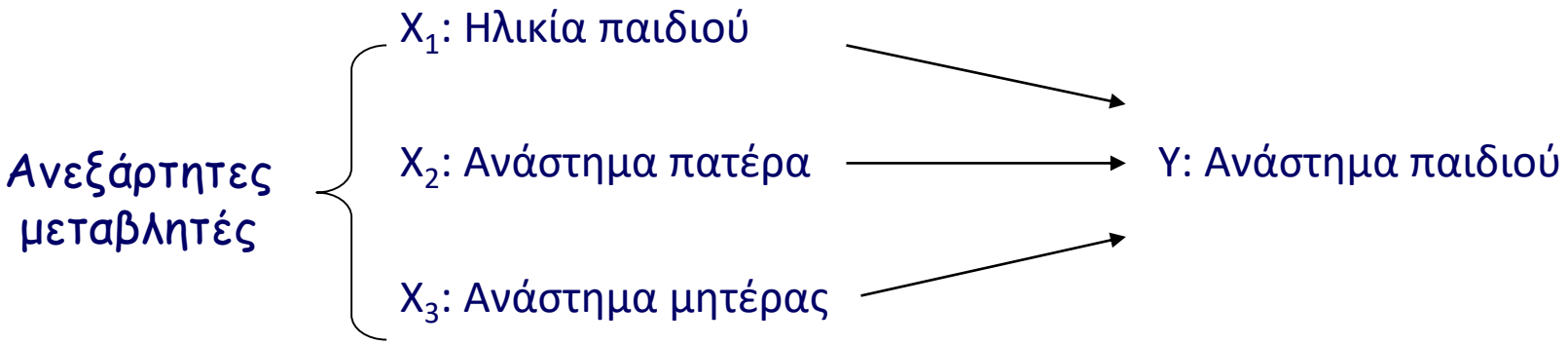
Οπου υπάρχει . υποδεικνύει ελλείπουσα τιμή. Για την πόλη: 1 σημαίνει Λουτράκι, 2 Λαύριο και 3 Ελευσίνα.

Κατανομή συχνοτήτων και Διάγραμμα συσχέτισης (στικτόγραμμα)





- Ας υποθέσουμε ότι θέλουμε να διερευνήσουμε την εξάρτηση του αναστήματος παιδιού (Υ) από την ηλικία του (X_1), το ανάστημα του πατέρα του (X_2) και από το ανάστημα της μητέρας του (X_3)



- Εκτιμούμε την ευθεία πολλαπλής γραμμικής εξάρτησης:

$$\hat{Y} = 20,50 + 4.56 * X_1 + 0.20 * X_2 + 0.21 * X_3$$

Ανάστημα παιδιού — Ηλικία παιδιού — Ανάστημα πατέρα — Ανάστημα μητέρας

Παράδειγμα πολλαπλής εξάρτησης για ανάστημα παιδιού (δείγμα 500 παιδιών)

Ανεξάρτητη μεταβλητή	b_i	SE_{bi}
Ηλικία παιδιού (έτη)	4,56	0,25
Ανάστημα πατέρα (cm)	0,20	0,03
Ανάστημα μητέρας (cm)	0,21	0,03

Ερμηνεία μερικού συντελεστή εξάρτησης της ηλικίας του παιδιού:

«Για κάθε ένα έτος αύξησης της ηλικίας του παιδιού παρατηρείται **μέση** αύξηση του αναστήματος κατά 4.56 cm, **ανεξάρτητα από το ανάστημα των γονέων» ή**

«Ένα παιδί που είναι ένα έτος μεγαλύτερο από ένα άλλο και οι γονείς τους έχουν ίδιο ανάστημα, αναμένεται να είναι 4,56 cm ψηλότερο **κατά μέσο όρο**»

Ερμηνεία μερικού συντελεστή εξάρτησης του αναστήματος του πατέρα:

Για κάθε ένα εκατοστό αύξησης του ύψους του πατέρα παρατηρείται **μέση** αύξηση του αναστήματος του παιδιού κατά 0,20 cm, ανεξάρτητα από το ανάστημα της μητέρας και της ηλικίας του παιδιού»

Ερμηνεία μερικού συντελεστή εξάρτησης του αναστήματος του μητέρας:

Για κάθε ένα εκατοστό αύξησης του ύψους της μητέρας παρατηρείται **μέση** αύξηση του αναστήματος του παιδιού κατά 0,21 cm, ανεξάρτητα από το ανάστημα του πατέρας και της ηλικίας του παιδιού»

ΕΙΝΑΙ ΑΥΤΕΣ ΟΙ ΣΧΕΣΕΙΣ ΣΤΑΤΙΣΤΙΚΑ ΣΗΜΑΝΤΙΚΕΣ;

Ανεξάρτητη μεταβλητή	b_i	SE_{bi}	t-test (b_i/SE_{bi})	B.E	P
Ηλικία παιδιού (έτη)	4,56	0,25	17,93	522-4 =518	$<10^{-6}$
Ανάστημα πατέρα (cm)	0,20	0,03	6,76	518	$<10^{-6}$
Ανάστημα μητέρας (cm)	0,21	0,03	6,13	518	$<10^{-6}$

Και οι 3 μεταβλητές είναι στατιστικά σημαντικές (p-value <0,05)

Άρα υπάρχει στατιστικά σημαντική σχέση του αναστήματος του παιδιού με την ηλικία και με το ανάστημα των γονέων.

Όρια αξιοπιστίας συντελεστών μερικής εξάρτησης

Τα **95% CI** υπολογίζονται από τον τύπο:

$$b_i \pm t_{0.05,(n-p-1)} SE(b_i)$$

Όπου **n**:μέγεθος δείγματος και **p**: αριθμός ΑΝΕΞΑΡΤΗΤΩΝ μεταβλητών

Π.χ. στο παράδειγμα, 95% CI για το b_i της ηλικίας παιδιού

$$b_i \pm t_{0.05,(n-p-1)} SE(b_i)$$

$$4{,}56 \pm 1{,}96 * 0{,}25$$

4,07 5,05

Με 95% πιθανότητα η μέση αύξηση του αναστήματος για αύξηση της ηλικίας του παιδιού κατά ένα έτος, και ανεξαρτήτως του ύψους των γονέων, βρίσκεται μεταξύ 4,07 και 5,05 cm.

Το 0 ΔΕΝ περιλαμβάνεται στο 95% ΔΕ του μερικού συντελεστή εξάρτησης: ΣΤΑΤΙΣΤΙΚΑ ΣΗΜΑΝΤΙΚΗ ΣΧΕΣΗ ηλικίας (X_1) και αναστήματος παιδιού (Y)

Γιατί είναι χρήσιμη μία τέτοια μεθοδολογία;

- Διερεύνηση παραγόντων που σχετίζονται με τα επίπεδα χοληστερόλης στα άτομα → ποιες παράμετροι εξηγούν τη μεταβλητότητα των επιπέδων χοληστερόλης μεταξύ των ατόμων;

  - Βαθύτερη κατανόηση των μηχανισμών
  - Αν βρεθεί ένα μοντέλο που εξηγεί ικανοποιητικά τη μεταβλητότητα, μπορεί να χρησιμοποιηθεί για πρόβλεψη
  - Αν εντοπιστούν παράγοντες που σχετίζονται με τα επίπεδα χοληστερόλης οι οποίοι μπορούν να τροποποιηθούν (π.χ. διατροφή, βάρος κλπ), τα αποτελέσματα μπορεί να αποτελέσουν βάση για συστάσεις-παρεμβάσεις

Ποιοτικές ανεξάρτητες μεταβλητές

- Σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης, οι ανεξάρτητες μεταβλητές μπορεί να είναι
  - Ποσοτικές
  - Ποιοτικές

  Π.χ. πως εξαρτάται το ανάστημα παιδιού από την ηλικία του (ποσοτική), το φύλο του (ποιοτική με 2 επίπεδα) και το επάγγελμα του πατέρα (ποιοτική με 3 επίπεδα)

- Πως εισάγονται οι ποιοτικές μεταβλητές στο μοντέλο και πως ερμηνεύονται οι συντελεστές μερικής εξάρτησης;

Ποιοτικές μεταβλητές με 2 επίπεδα

Αν η ποιοτική μεταβλητή έχει 2 επίπεδα (π.χ. φύλο):

- Αποφασίζουμε ποια σύγκριση επιθυμούμε να κάνουμε π.χ. άνδρες σε σχέση με γυναίκες ή αντίστροφα
  - Π.χ. αγόρια σε σχέση με κορίτσια →

    κατηγορία αναφοράς: κορίτσια

- Η κατηγορία αναφοράς κωδικοποιείται με 0 και η άλλη κατηγορία με 1 (γενικά η κατηγορία αναφοράς κωδικοποιείται με τη μικρότερη τιμή)

- Ο **συντελεστής b** για το φύλο εκφράζει πόσο διαφέρουν κατά μέσο όρο τα αγόρια από τα κορίτσια ως προς το ανάστημά τους (κατηγορία αναφοράς: κορίτσια), ενώ η **σταθερά a** τη μέση τιμή της εξαρτημένης για τα άτομα της ΚΑΤΗΓΟΡΙΑΣ ΑΝΑΦΟΡΑΣ (κορίτσια).

Ποιοτικές μεταβλητές με >2 επίπεδα

- Μια ποιοτική μεταβλητή με περισσότερα των δύο επιπέδων απαιτείται η δημιουργία ψευδομεταβλητών (dummy variables or indicator variables).

  - π.χ. εκπαίδευση κωδικοποιημένο ως:
    - Απόφοιτος Γυμνασίου: 1
    - Απόφοιτος Λυκείου:2
    - Πανεπιστημιακής εκπαίδευσης:3

Ποιοτικές μεταβλητές με >2 επίπεδα

Δημιουργία ψευδομεταβλητών (dummy / indicator variables)

Παράδειγμα: εκπαίδευση

1. Απόφοιτος Γυμνασίου -> ψευδομεταβλητή X_1

= 1 αν απόφοιτος γυμνασίου

= 0 άλλο

2. Απόφοιτος Λυκείου -> ψευδομεταβλητή X_2

= 1 αν απόφοιτος Λυκείου

= 0 άλλο

3. Πανεπιστημιακής μόρφωσης -> ψευδομεταβλητή X_3

= 1 αν έχει πανεπιστημιακή μόρφ.

= 0 άλλο

Πώς εισάγουμε ποιοτικές μεταβλητές;

- Στο μοντέλο εισάγονται **Κ-1** ψευδομεταβλητές, όπου Κ ο αριθμός των επιπέδων της ποιοτικής μεταβλητής.
  - Αυτή που **δεν** εισέρχεται: κατηγορία αναφοράς (reference category)

  - Εκτιμώνται τα b_i για τη σύγκριση κάθε μίας κατηγορίας προς την κατηγορία αναφοράς, π.χ.
    - Απ. Λυκείου (ψευδομεταβλητή X_2)σε σχέση με Γυμνασίου→ ο συντελεστής b της X_2
    - πανεπιστ. μόρφωσης (ψευδομεταβλητή X_3) σε σχέση με Γυμνασίου→ ο συντελεστής b της X_3

  - Ο συντελεστής μερικής εξάρτησης : η μέση διαφορά στην εξαρτημένη μεταβλητή για άτομα της κατηγορίας που αντιστοιχεί η ψευδομεταβλητή **σε σχέση με άτομα στην κατηγορία αναφοράς**.

Στο προηγούμενο μοντέλο είχαμε:
Εξαρτημένη μεταβλητή: ανάστημα παιδιού(cm)
Ανεξάρτητες μεταβλητές: ηλικία παιδιού (έτη), ανάστημα πατέρα (cm), ανάστημα μητέρας (cm) -> ποσοτικές μεταβλητές

Ανεξάρτητη μεταβλητή	b_i	SE_{bi}	t-test (b_i/SE_{bi})	B.E	P
Ηλικία παιδιού (έτη)	4,56	0,25	17,93	522-4 =518	$<10^{-6}$
Ανάστημα πατέρα (cm)	0,20	0,03	6,76	518	$<10^{-6}$
Ανάστημα μητέρας (cm)	0,21	0,03	6,13	518	$<10^{-6}$

Και οι 3 μεταβλητές είναι στατιστικά σημαντικές

Στο επόμενο μοντέλο έχουμε :
Εξαρτημένη μεταβλητή: ανάστημα παιδιού(cm)
Ανεξάρτητες μεταβλητές: φύλο παιδιού, επάγγελμα πατέρα -> ποιοτικές μεταβλητές

Ερμηνεία ψευδομεταβλητών

Ύψος παιδιού (Υ) σε σχέση με επάγγελμα πατέρα

Παράγοντας	b	SE(b)	p-value
Επάγγελμα πατέρα			
Πανεπ. Μόρφωσης	Κατηγορία αναφοράς		
Ανειδίκευτος	-2.41	0.89	<0.05
Ειδικευμένος	+0.17	0.83	>0.10
Φύλο			
Αγόρι	Κατηγορία αναφοράς		
Κορίτσι	-0.66	0.44	>0.10

- **Ερμηνεία συντελεστών της εκπαίδευσης του πατέρα:**

  Κατηγορία αναφοράς: πανεπ. εκπαίδευση

  - **Ανεξαρτήτως του φύλου τους,** τα παιδιά των ανειδίκευτων, έχουν **χαμηλότερο** ανάστημα από τα παιδιά αυτών με πανεπιστημιακή μόρφωση (κατηγορία αναφοράς) κατά μέσο όρο κατά 2,41 cm, και η διαφορά αυτή είναι στατιστικά σημαντική (p-value<0,05).
  - **Ανεξαρτήτως του φύλου τους,** τα παιδιά των ειδικευμένων, έχουν **υψηλότερο** ανάστημα από τα παιδιά αυτών με πανεπιστημιακή μόρφωση (κατηγορία αναφοράς) κατά μέσο όρο κατά 0,17 cm αλλά η διαφορά αυτή δεν είναι στατιστικά σημαντική (p-value>0,05) (Το ύψος των παιδιών των ειδικευμένων **ΔΕΝ ΔΙΑΦΕΡΕΙ** σε βαθμό στατιστικά σημαντικό από το ύψος των παιδιών με πατέρα πανεπιστημιακής μόρφωσης.)

- **Ερμηνεία συντελεστή του φύλου:**

  Κατηγορία αναφοράς: αγόρια

  Ανεξαρτήτως του επαγγέλματος του πατέρα, τα κορίτσια έχουν **χαμηλότερο** ύψος από τα αγόρια (κατηγορία αναφοράς) κατά μέσο όρο κατά 0,66 cm **αλλά η διαφορά δεν είναι στατιστικά σημαντική (p-value > 0,05)**

Αποτελέσματα πολλαπλής γραμμικής εξάρτησης για τη σχέση μεταξύ βάρους νεογνού με άλλους παράγοντες

95% διάστημα αξιοπιστίας

Στήλη με p-value

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1755.228	1150.652		-1.525	.129	-4023.535	513.080
	gestdur	125.486	27.975	.294	4.486	.000	70.338	180.634
	gender	-94.722	65.854	-.095	-1.438	.152	-224.542	35.098
	age	10.127	7.341	.091	1.380	.169	-4.344	24.598

a. Dependent Variable: bweight

Στήλη μερικών συντελεστών εξάρτησης

Ερμηνεία συντελεστών ηλικίας (age) και διάρκειας κύησης (gestdur)-> <u>ποσοτικές μεταβλητές:</u>

1. Ανεξαρτήτως φύλου του νεογνού και της ηλικίας της μητέρας, για αύξηση της διάρκειας κύησης κατά 1 εβδομάδα, το βάρος του νεογνού αναμένεται **να αυξηθεί κατά μέσο όρο** κατά 125,5 γραμμάρια (με 95% πιθανότητα η μεταβολή αυτή να κυμαίνεται από 70,3 έως 180,6 γραμμάρια) -> στατιστικά σημαντική σχέση (p-value <0,05 και το 95% διάστημα αξιοπιστίας δεν περιλαμβάνει το 0)
2. Ανεξαρτήτως φύλου του νεογνού και διάρκειας κύησης, για αύξηση της ηλικίας της μητέρας κατά 1 έτος, το βάρος του νεογνού αναμένεται **να αυξηθεί κατά μέσο όρο** κατά 10,1 γραμμάρια (με 95% πιθανότητα η μεταβολή αυτή να κυμαίνεται από -4,3 έως 24,6 γραμμάρια) -> μη στατιστικά σημαντική σχέση (p-value >0,05 και το 95% διάστημα αξιοπιστίας περιλαμβάνει το 0)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1755.228	1150.652		-1.525	.129	-4023.535	513.080
	gestdur	125.486	27.975	.294	4.486	.000	70.338	180.634
	gender	-94.722	65.854	-.095	-1.438	.152	-224.542	35.098
	age	10.127	7.341	.091	1.380	.169	-4.344	24.598

a. Dependent Variable: bweight

Ερμηνεία συντελεστή του φύλλου του νεογνού (gender)-> <u>ποιοτική μεταβλητή:</u>

Κατηγορία αναφοράς: τα αγόρια (δε φαίνεται στον πίνακα με τους συντελεστές του SPSS)

Ανεξαρτήτως της ηλικίας της μητέρας και της διάρκειας κύησης, το βάρος των κοριτσιών αναμένεται να είναι **κατά μέσο 94,7 γραμμάρια χαμηλότερο** από τα αγόρια (**κατηγορία αναφοράς**) (με 95% πιθανότητα η διαφορά αυτή να κυμαίνεται από -224,5 έως 35,1 γραμμάρια) -> μη στατιστικά σημαντική σχέση (p-value >0,05 και το 95% διάστημα αξιοπιστίας περιλαμβάνει το 0)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.372[a]	.138	.116	463.84034

a. Predictors: (Constant), edu3, gender, gestdur, age, edu2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-2585.409	1203.890		-2.148	.033	-4959.803	-211.015
	gestdur	139.855	28.884	.327	4.842	.000	82.889	196.822
	gender	-121.563	67.495	-.121	-1.801	.073	-254.682	11.555
	age	20.317	9.734	.180	2.087	.038	1.120	39.514
	edu2	38.225	89.635	.039	.426	.670	-138.560	215.009
	edu3	-186.018	114.652	-.165	-1.622	.106	-412.142	40.107

a. Dependent Variable: bweight

Έχει προστεθεί στο προηγούμενο μοντέλο μία ακόμα ποιοτική μεταβλητή (επίπεδο εκπαίδευσης μητέρας) με 3 επίπεδα:

1) Χαμηλό επίπεδο (κατηγορία αναφοράς-η αντίστοιχη ψευδομεταβλητή δεν εισέρχεται στο μοντέλο)
2) Μέσο επίπεδο (αντίστοιχη ψευδομεταβλητή edu2)
3) Υψηλό επίπεδο (αντίστοιχη ψευδομεταβλητή edu3)

Ερμηνεία συντελεστών ηλικίας (age) και διάρκειας κύησης (gestdur)-> <u>ποσοτικές μεταβλητές:</u>

1. Ανεξαρτήτως φύλλου του νεογνού, της ηλικίας της μητέρας και της εκπαίδευσης της μητέρας, για αύξηση της διάρκειας κύησης κατά 1 εβδομάδα, το βάρος του νεογνού αναμένεται **να αυξηθεί κατά μέσο όρο** κατά 139,9 γραμμάρια (με 95% πιθανότητα η μεταβολή αυτή να κυμαίνεται από 82,9 έως 196,8 γραμμάρια) -> στατιστικά σημαντική σχέση (p-value <0,05 και το 95% διάστημα αξιοπιστίας δεν περιλαμβάνει το 0)

2. Ανεξαρτήτως φύλου του νεογνού, της εκπαίδευσης της μητέρας και της διάρκειας κύησης, για αύξηση της ηλικίας της μητέρας κατά 1 έτος, το βάρος του νεογνού αναμένεται **να αυξηθεί κατά μέσο όρο** κατά 20,3 γραμμάρια (με 95% πιθανότητα η μεταβολή αυτή να κυμαίνεται από 1,1 έως 39,5 γραμμάρια) -> στατιστικά σημαντική σχέση (p-value <0,05 και το 95% διάστημα αξιοπιστίας δεν περιλαμβάνει το 0)

Ερμηνεία συντελεστών φύλου (gender) και εκπαίδευσης μητέρας (edu2,edu3)-> <u>ποιοτικές μεταβλητές:</u>

1. Κατηγορία αναφοράς για το φύλο: τα αγόρια (δε φαίνεται στον πίνακα με τους συντελεστές του SPSS)

Ανεξαρτήτως της ηλικίας της μητέρας, της διάρκειας κύησης και της εκπαίδευσης της μητέρας, το βάρος των κοριτσιών αναμένεται να είναι **κατά μέσο όρο 121,6 γραμμάρια χαμηλότερο** από τα αγόρια (**κατηγορία αναφοράς**) (με 95% πιθανότητα η διαφορά αυτή να κυμαίνεται από -254,7 έως 11,6 γραμμάρια) -> μη στατιστικά σημαντική σχέση (p-value >0,05 και το 95% διάστημα αξιοπιστίας περιλαμβάνει το 0)

2. Κατηγορία αναφοράς για το επίπεδο εκπαίδευσης: το χαμηλό επίπεδο (δε φαίνεται στον πίνακα με τους συντελεστές του SPSS)

- Ανεξαρτήτως των υπόλοιπων παραγόντων, το βάρος των νεογνών των οποίων η μητέρα έχει μέσο επίπεδο εκπαίδευσης (edu2) αναμένεται να είναι **κατά μέσο όρο 38,2 γραμμάρια υψηλότερο** από το βάρος των νεογνών των οποίων η μητέρα έχει χαμηλό επίπεδο εκπαίδευσης(**κατηγορία αναφοράς**) (με 95% πιθανότητα η διαφορά αυτή να κυμαίνεται από -138,6 έως 215 γραμμάρια) -> μη στατιστικά σημαντική σχέση (p-value >0,05 και το 95% διάστημα αξιοπιστίας περιλαμβάνει το 0).

- Ανεξαρτήτως των υπόλοιπων παραγόντων, το βάρος των νεογνών των οποίων η μητέρα έχει υψηλό επίπεδο εκπαίδευσης (edu3) αναμένεται να είναι **κατά μέσο όρο 186 γραμμάρια χαμηλότερο** από το βάρος των νεογνών των οποίων η μητέρα έχει χαμηλό επίπεδο εκπαίδευσης(**κατηγορία αναφοράς**) (με 95% πιθανότητα η διαφορά αυτή να κυμαίνεται από -412,1 έως 40,1 γραμμάρια) -> μη στατιστικά σημαντική σχέση (p-value >0,05 και το 95% διάστημα αξιοπιστίας περιλαμβάνει το 0).

Επιλογή των ανεξάρτητων μεταβλητών

Α. Λόγοι επιλογής

- Να έχει ιδιαίτερο ενδιαφέρον από μόνη της
- Να αποτελεί σημαντικό προγνωστικό παράγοντα
- Τον έλεγχο των πιθανών συγχυτικών επιδράσεων

Β. Τρόποι επιλογή

Ο αριθμός των παραμέτρων (*ανεξάρτητες μεταβλητές + σταθερά του μοντέλου*) πρέπει να είναι σαφώς μικρότερος του αριθμού των παρατηρήσεων *n* (το πολύ ίσος με *n*/10)

Το τελικό μοντέλο να είναι επεξηγηματικό και ταυτόχρονα λιτό

Συντελεστής πολλαπλής συσχέτισης R^2

- R^2: Το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από τις ανεξάρτητες μεταβλητές (μπορεί να αποδοθεί στις ανεξάρτητες μεταβλητές)

  - Τιμές από **0** (το μοντέλο δεν εξηγεί καθόλου τη μεταβλητότητα) **έως 1** (το μοντέλο εξηγεί 100% τη μεταβλητότητα)
  - R^2 αυξάνεται κάθε φορά που προστίθεται μία μεταβλητή στο μοντέλο, ανεξάρτητα από το πόσο σημαντική είναι αυτή η μεταβλητή → Προσαρμοσμένο (adjusted) R^2: **ΔΕΝ ΑΥΞΑΝΕΙ** πάντα με τη προσθήκη νέας μεταβλητής στο μοντέλο

Όταν συγκρίνω μοντέλα με διαφορετικό αριθμό ανεξάρτητων μεταβλητών, καλύτερο μοντέλο είναι αυτό με ***το μεγαλύτερο προσαρμοσμένο*** *R^2 (οι μεταβλητές αυτές εξηγούν μεγαλύτερο ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής)*

Συνοψίζοντας

Απλή γραμμική εξάρτηση 2 <u>ποσοτικών</u> μεταβλητών

- Γραμμική σχέση
- Εξάρτηση της Υ από την Χ
- Μόνο η Υ -> κανονική κατανομή

<u>Συντελεστής εξάρτησης b</u>:

Κατά μέσο όρο μεταβολή της Υ , για κάθε μία μονάδα αύξησης της Χ

Στατιστική αξιολόγηση- Είναι στατιστικά σημαντικό το εύρημα;

H_0: απουσία εξάρτησης (β=0) στον πληθυσμό

H_1: παρουσία εξάρτησης (β≠0) στον πληθυσμό

Πολλαπλή γραμμική εξάρτηση μιας ποσοτικής με παραπάνω από μία ποσοτικές ή ποιοτικές μεταβλητές

- Γραμμική σχέση
- Εξάρτηση της Y από τις X1,X2,…,Xp
- Μόνο η Y -> κανονική κατανομή

Συντελεστές μερικής εξάρτησης bi ποσοτικών μεταβλητών :

- Κατά μέσο όρο μεταβολή της Y , για κάθε μία μονάδα αύξησης της Xi, όταν όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές (δηλαδή ανεξάρτητα από τις άλλες ανεξάρτητες μεταβλητές)

Συντελεστές μερικής εξάρτησης bi ποιοτικών μεταβλητών :

- Κατά μέσο όρο διαφορά της Y, των ατόμων της μιας κατηγορίας σε σχέση με τα άτομα της κατηγορίας αναφοράς, όταν όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές (δηλαδή ανεξάρτητα από τις άλλες ανεξάρτητες μεταβλητές)

Στατιστική αξιολόγηση- Είναι στατιστικά σημαντικό το εύρημα

H_0: απουσία εξάρτησης ($\beta i=0$)
H_1: παρουσία εξάρτησης ($\beta i\neq 0$)

Συντελεστής πολλαπλής συσχέτισης R^2

- Στα μοντέλα γραμμικής εξάρτησης, μπορώ να υπολογίσω και τον συντελεστή συσχέτισης **R^2** ο οποίος μας δείχνει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής Υ που ερμηνεύεται από τις ανεξάρτητες μεταβλητές Χ που έχουν εισαχθεί στο μοντέλο.

- Σύγκριση μοντέλων με προσαρμοσμένο **R^2 :**
  Μεγαλύτερος **R^2** => Καλύτερο μοντέλο