Δ. ΣΤΑΤΙΣΤΙΚΗ ΣΥΣΧΕΤΙΣΗ

Δ_1. Υπολογισμός συντελεστών συσχέτισης

Προκειμένου να ελέγξουμε την ύπαρξη γραμμικής σχέσης μεταξύ δύο ποσοτικών μεταβλητών, χρησιμοποιούμε συνήθως τον παραμετρικό συντελεστή συσχέτισης του **Pearson**, r.

Προϋποθέσεις :

- Και οι δύο μεταβλητές να κατανέμονται κανονικά και να έχουν επιλεγεί τυχαία.

Στην περίπτωση που δεν ισχύει η προϋπόθεση της κανονικότητας των μεταβλητών, υπολογίζουμε τον αντίστοιχο μη παραμετρικό συντελεστή του **Spearman, r_s**.

Ιδιότητες των συντελεστών:

- Είναι καθαροί αριθμοί (δεν έχουν μονάδες).
- Παίρνουν τιμές από -1 (*αρνητική συσχέτιση*), έως +1 (*θετική συσχέτιση*). Όπου:
  - *Αρνητική συσχέτιση*: μικρές τιμές της μίας μεταβλητής αντιστοιχούν σε μεγάλες τιμές της άλλης και αντίστροφα..
  - *Θετική συσχέτιση*: μικρές τιμές της μίας μεταβλητής αντιστοιχούν σε μικρές τιμές της άλλης και αντίστροφα..

  Δηλαδή, το πρόσημο των συντελεστών καταδεικνύει το είδος της σχέσης, ενώ όσο μεγαλύτερη είναι η απόλυτη τιμή τους, τόσο ισχυρότερη είναι η συσχέτιση των δύο μεταβλητών. Τέλος, η τιμή μηδέν αντιστοιχεί στη μη ύπαρξη γραμμικής σχέσης.
- Είναι μέτρα του βαθμού της γραμμικής σχέσης.

Μηδενική υπόθεση: Ο συντελεστής συσχέτισης είναι ίσος με το μηδέν (r ή $r_s = 0$, οι μεταβλητές δεν σχετίζονται).

έναντι της

Εναλλακτικής υπόθεσης: Ο συντελεστής συσχέτισης είναι διάφορος του μηδενός (r ή $r_s \neq 0$, οι μεταβλητές σχετίζονται).

SPSS

1) *Παραμετρικό τεστ*

Για παράδειγμα, ας υποθέσουμε ότι θέλουμε να ελέγξουμε, αν υπάρχει σχέση ανάμεσα στο βάρος των γυναικών, πριν την δίαιτα, και στο ύψος τους (είναι και οι δύο κανονικές μεταβλητές):

(Αφού έχει προηγηθεί έλεγχος κανονικότητας Kolmogorov – Smirnov, βλ. Class_2)

Analyze → Correlate → Bivariate → Variables: βάζουμε τις μεταβλητές που θέλουμε να ελέγξουμε την συσχέτισή τους, **Correlation coefficient**: Pearson **→ Ok**

Correlations

		Ύψος	Βάρος (πριν)
Ύψος	Pearson Correlation	1,000	,291**
	Sig. (2-tailed)	,	,000
	N	162	161
Βάρος (πριν)	Pearson Correlation	,291**	1,000
	Sig. (2-tailed)	,000	,
	N	161	174

**. Correlation is significant at the 0.01 level (2-tailed).

σχόλια:

Ο r είναι ίσος με +0.291, γεγονός που καταδεικνύει θετική γραμμική συσχέτιση. Αυτό σημαίνει ότι μικρές τιμές του ύψους αντιστοιχούν σε μικρές τιμές του βάρους και αντίστροφα. Το r δεν είναι πολύ κοντά στο ένα, δηλαδή δεν έχουμε συσχέτιση ιδιαίτερα μεγάλου βαθμού. Ωστόσο, είναι στατιστικά πολύ σημαντική, στο επίπεδο του 1 % (p-value < 0.0001).

2) Μη - Παραμετρικό τεστ

Ας υποθέσουμε ότι θέλουμε να ελέγξουμε, αν υπάρχει σχέση ανάμεσα στην ηλικία των γυναικών (κατανέμεται κανονικά) και στα έτη σπουδών (δεν κατανέμεται κανονικά) :

Analyze → Correlate → Bivariate → Variables: βάζουμε τις μεταβλητές που θέλουμε να ελέγξουμε την συσχέτισή τους, **Correlation coefficient**: Spearman → **Ok**

Correlations

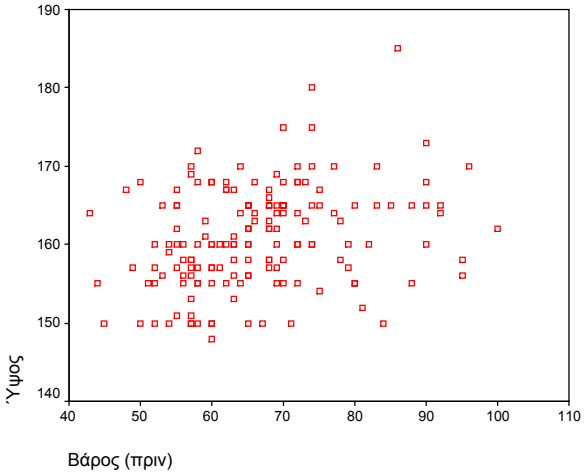
			Έτη σχολικής φοίτησης	Ηλικία
Spearman's rho	Έτη σχολικής φοίτησης	Correlation Coefficient	1,000	-,382**
		Sig. (2-tailed)	,	,000
		N	175	175
	Ηλικία	Correlation Coefficient	**-,382****	1,000
		Sig. (2-tailed)	**,000**	,
		N	175	175

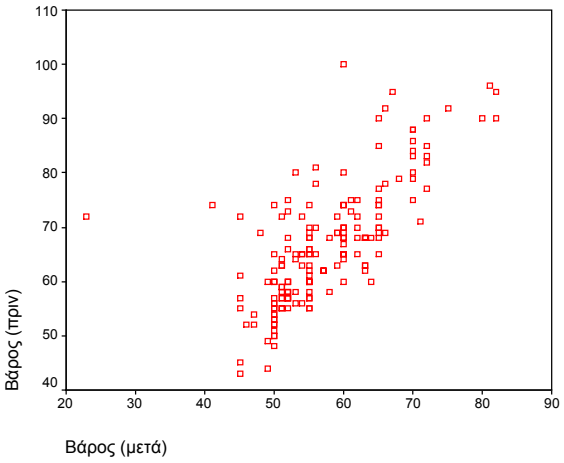
**. Correlation is significant at the .01 level (2-tailed).

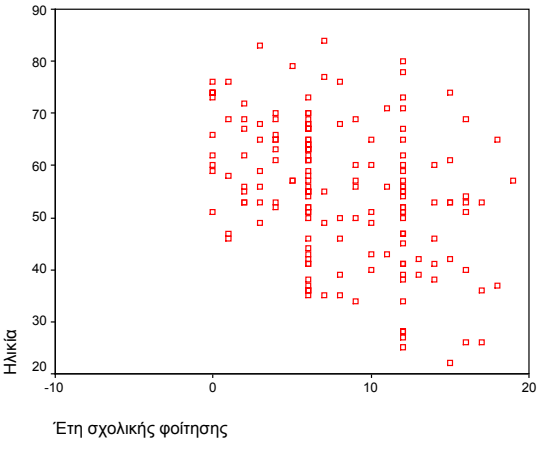
Ο r_s είναι ίσος με -0.382, γεγονός που καταδεικνύει αρνητική γραμμική συσχέτιση. Αυτό σημαίνει ότι μεγάλες ηλικίες αντιστοιχούν σε λίγα χρόνια σπουδών και αντίστροφα Η απόλυτη τιμή του r_s δεν είναι πολύ κοντά στο ένα, δηλαδή δεν φαίνεται να υπάρχει συσχέτιση μεγάλου βαθμού. Ωστόσο είναι στατιστικά πολύ σημαντική, στο επίπεδο του 1 % (p-value < 0.0001).

Δ_2. Έλεγχος συσχέτισης ποσοτικών μεταβλητών γραφικά.

Graphs → Scatter Plot → Simple→ Define: Τοποθετούμε τις μεταβλητές στους δύο άξονες → **Ok**







Ε. ΣΤΑΤΙΣΤΙΚΗ ΕΞΑΡΤΗΣΗ – ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

E_1. Απλή γραμμική παλινδρόμηση

Όπως είδαμε, ο συντελεστής συσχέτισης μας πληροφορεί για το **αν** και κατά **πόσο** δύο μεταβλητές σχετίζονται. Ωστόσο δεν μας πληροφορεί για το **πως** σχετίζονται. Δηλαδή, τον τρόπο με τον οποίο μεταβάλλονται οι τιμές τις μίας, συναρτήσει της άλλης. Σε αυτή την περίπτωση, δηλαδή όταν θέλουμε *να διερευνήσουμε τη μεταβολή των τιμών της μίας μεταβλητής (εξαρτημένη), συναρτήσει των μεταβολών της άλλης (ανεξάρτητης)* εφαρμόζουμε στατιστική εξάρτηση ή παλινδρόμηση.

Προϋποθέσεις :

- Η εξαρτημένη μεταβλητή να κατανέμεται κανονικά και να έχει επιλεγεί τυχαία.

Μοντέλο απλής γραμμικής παλινδρόμησης :

Η κύρια ιδέα της γραμμικής εξάρτησης, είναι η δημιουργία μίας ευθείας, που να εφαρμόζει καλύτερα στα δεδομένα. Η ευθεία αυτή περιγράφεται από την εξίσωση:

$$E(\hat{Y}_i \mid X_i) = b_0 + b_1 X_i \qquad (1)$$

όπου b_0 είναι η σταθερά της και b_1 η κλίση της. Δηλαδή, η μέση τιμή της εξαρτημένης μεταβλητής Υ μεταβάλλεται με σταθερό ρυθμό, όταν μεταβάλλονται οι τιμές της ανεξάρτητης μεταβλητής. Έτσι για κάθε μία παρατήρηση της Υ, εκτιμάμε :

$$\hat{Y}_i = b_0 + b_1 X_i + e_i \qquad (2)$$

όπου τα e_i είναι τα παρατηρηθέντα υπόλοιπα, δηλαδή η διαφορά της τιμής της εξαρτημένης μεταβλητής Υ στο δείγμα, για δεδομένο Χ (παρατηρούμενη τιμή του Υ),

από την τιμή που αναμένουμε για την Y με βάση την εξίσωση (1) (βλ. Γραφική απεικόνιση).

Συντελεστής εξάρτησης :

Η κλίση της ευθείας, δηλαδή το b_1, καλείται συντελεστής εξάρτησης, για τον οποίο ισχύουν:

- Διαθέτει μονάδες, το λόγο των μονάδων της εξαρτημένης μεταβλητής προς τις μονάδες της ανεξάρτητης.
- Μπορεί να πάρει οποιαδήποτε τιμή. Ο συντελεστής εξάρτησης μπορεί να είναι αρνητικός (αρνητική εξάρτηση), θετικός (θετική εξάρτηση) ενώ όταν είναι μηδέν δεν υπάρχει εξάρτηση μεταξύ των υπό εξέταση μεταβλητών.
- Η μεταβολή της εξαρτημένης, συναρτήσει της μεταβολής στην ανεξάρτητη, δίνεται μέσο του b_1 κατά τον παρακάτω τρόπο.

Ας θεωρήσουμε το μοντέλο:

$$E(\hat{Y}_i \mid X_i) = b_0 + b_1 \cdot X_i \qquad (3)$$

Αν υποθέσουμε ότι η X αυξάνεται κατά μία μονάδα, η 3 γράφεται :

$$E(\hat{Y}_i \mid (X_i + 1)) = b_0 + b_1 \cdot (X_i + 1)$$

(4)

Προκειμένου να υπολογίσουμε την μεταβολή στην μέση τιμή της Y, για μία μονάδα αύξησης της X, αφαιρούμε τις παραπάνω σχέσεις.

$$E(\hat{Y}_i \mid X_i) - E(\hat{Y}_i \mid (X_i + 1)) = b_0 + b_1 \cdot (X_i + 1) - b_0 - b_1 \cdot X_i = b_1 \qquad (5)$$

Είναι εμφανές ότι ο συντελεστής εξάρτησης b_1 εκφράζει το μέσο όρο της μεταβολής της εξαρτημένης μεταβλητής, όταν η ανεξάρτητη μεταβληθεί κατά μία μονάδα. Τιμή του b_1 ίση με το μηδέν, κατ' επέκταση, ισοδυναμεί με απουσία εξάρτησης.

Προϋποθέσεις καταλοίπων για την εφαρμογή γραμμικής εξάρτησης :

- Η μέση τιμή των πραγματικών υπολοίπων ε_i να είναι 0.
- Η διακύμανση των ε_i να είναι σταθερή.
- Τα ε_i να είναι ανά δύο ανεξάρτητα και να κατανέμονται κανονικά.

Οι παραπάνω προϋποθέσεις των καταλοίπων στον πληθυσμό, εκτιμούνται από τα αντίστοιχα δειγματικά κατάλοιπα e_i, μετά την εφαρμογή της μεθόδου.

SPSS

Ας υποθέσουμε για παράδειγμα, ότι θέλουμε να μελετήσουμε τον τρόπο με τον οποίο μεταβάλλεται το βάρος των γυναικών *(εξαρτημένη μεταβλητή),* σε σχέση με το ύψος τους *(ανεξάρτητη)*. Η μεταβλητή βάρος κατανέμεται κανονικά, οπότε μπορούμε να εφαρμόσουμε την μέθοδο.

Analyze → Regression → Linear → και βάζουμε την εξαρτημένη και την ανεξάρτητη μεταβλητή μας στο κατάλληλο πλαίσιο → Ok

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,291[a]	,084	,079	10,91

a. Predictors: (Constant), Ύψος

σχόλιο:

$\mathbf{R^2 = r^2}$, δηλαδή το τετράγωνο του συντελεστή συσχέτισης μεταξύ των μεταβλητών.

$\mathbf{R^2}$ = 0.084 → Το μοντέλο επεξηγεί το 8.4% της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής.

$\mathbf{R^2}$ **= Sum of Square (regression)/ Sum of Square (total)** (βλέπε επόμενο πίνακα).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1745,075	1	1745,075	14,658	,000[a]
	Residual	18929,099	159	119,051		
	Total	20674,174	160			

a. Predictors: (Constant), Ύψος

b. Dependent Variable: Βάρος (πριν)

σχόλιο:

Mean Square = Sum of Squares / df

M. S. (regression) = 1745.075 → η μεταβλητότητα που ερμηνεύεται από το μοντέλο.

M. S. (residuals) = 119.051 → η μεταβλητότητα που υπολείπεται, δηλαδή η μεταβλητότητα που οφείλεται σε άλλους παράγοντες, τους οποίους δεν έχουμε εισάγει στο μοντέλο μας.

F = M. S. (regression)/ M. S. (residuals) → ελέγχουμε αν το μοντέλο εξηγεί σημαντικό μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής. Ο έλεγχος γίνεται κάτω από την μηδενική υπόθεση: MS(reg)=MS(res), δηλαδή F=1.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-15,514	21,444		-,723	,470	-57,867	26,838
	Ύψος	,510	,133	,291	3,829	,000	,247	,773

a. Dependent Variable: Βάρος (πριν)

σχόλια:

Το μοντέλο που αντιστοιχεί στον παραπάνω πίνακα είναι το:

$$E(\hat{Y}_i \mid X_i) = -15.514 + 0.510 \cdot X_i$$

Δηλαδή : Constant = b_0 = -15.514 και b_1 = 0.51

Για κάθε έναν από τους συντελεστές, γίνεται ο έλεγχος one sample t - test, κάτω από την μηδενική υπόθεση H_0: *ο συντελεστής είναι ίσος με το μηδέν* (που ισοδυναμεί στη περίπτωση του b_i, με απουσία εξάρτησης).

b_1 = + 0.51 → για κάθε εκατοστό αύξηση του ύψους, αναμένεται κατά μέσο όρο αύξηση κατά 0.51 κιλά του βάρους. Το αποτέλεσμα αυτό είναι στατιστικά σημαντικό στο επίπεδο του 1% (p-value <0.0001).

b_0 = -15.514 → Η τιμή αυτή αντιστοιχεί στην αναμενόμενη μέση τιμή του βάρους, για ύψος ίσο με το μηδέν! Καταλαβαίνουμε βέβαια πως κάτι τέτοιο δεν έχει φυσική ερμηνεία. Σε αυτή την περίπτωση καταφεύγουμε σε στάθμιση (κεντράρισμα) της μεταβλητής, αφαιρώντας από κάθε παρατήρηση την μέση τιμή της.

SPSS

Transform → Compute → Target variable: το όνομα της νέας μεταβλητής, **Numeric Expression**: τη μεταβλητή μείον τη μέση τιμή της → **Ok**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66,488	,860		77,315	,000
	Σταθμισμένο Ύψος	,510	,133	,291	3,829	,000

a. Dependent Variable: Βάρος (πριν)

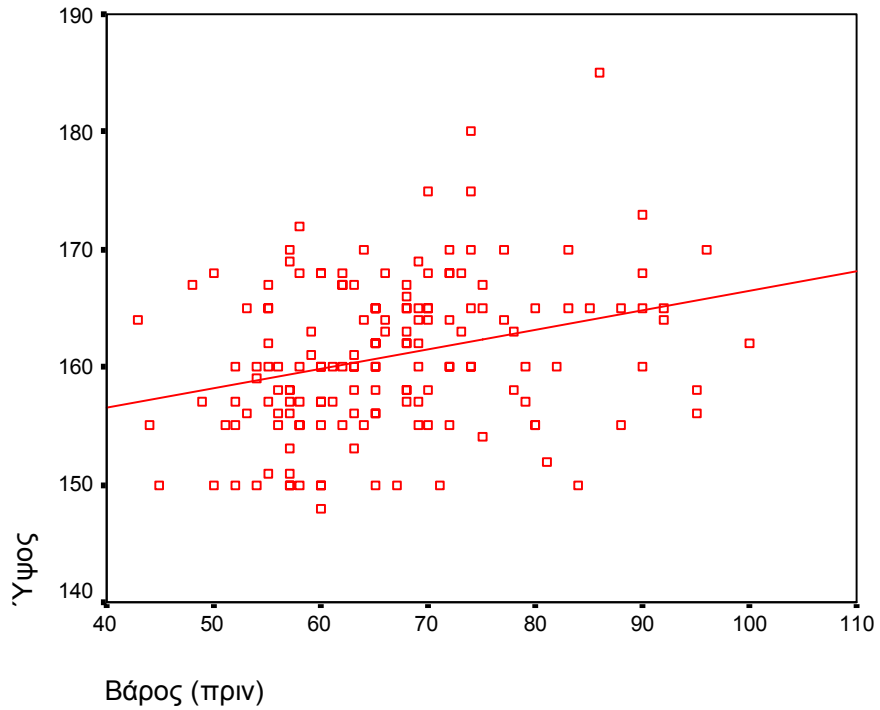
Το νέο μοντέλο που προκύπτει είναι: $E(\hat{Y}_i \mid X_i) = 66.488 + 0.510 \cdot (X_i - \bar{X})$

b_0 = 66.488 → Η τιμή αυτή αντιστοιχεί στο βάρος που αναμένουμε να έχουν οι γυναίκες που έχουν ύψος ίσο με την μέση τιμή, δηλαδή 160.84 cm.

Η τιμή και η ερμηνεία του b_1 δεν αλλάζει.

$E_{2.}$ Γραφική απεικόνιση

Graphs → Scatter Plot → Simple→ Define: Τοποθετούμε τις μεταβλητές στους δύο άξονες **→ Ok .** Με διπλό κλικ στο γράφημα οδηγούμαστε στο chart editor , όπου επιλέγουμε **Chart → Options → Fit line total,** προκειμένου να εισάγουμε στο στικτόγραμμα την ευθεία της εξάρτησης (γραμμή ελαχίστων τετραγώνων).



E_2. Πολλαπλή γραμμική παλινδρόμηση

Στην περίπτωση που θέλουμε *να διερευνήσουμε τη μεταβολή των τιμών μίας μεταβλητής (εξαρτημένη) συναρτήσει των τιμών όχι μίας αλλά περισσότερων ανεξάρτητων μεταβλητών,* εφαρμόζουμε πολλαπλή στατιστική εξάρτηση ή παλινδρόμηση.

Μοντέλο πολλαπλής γραμμικής παλινδρόμησης :

Το αντίστοιχο μοντέλο για τις μεταβολές της μέσης τιμής της εξαρτημένης μεταβλητής, συναρτήσει των τιμών των ανεξάρτητων είναι το:

$$E(\hat{Y}_i \mid X_{1i}, X_{2i}, ..., X_{pi}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi} \qquad (6)$$

Όσα αναφέρθηκαν στην προηγούμενη παράγραφο όσων αφορά τις προϋποθέσεις εφαρμογής της παλινδρόμησης, ισχύουν και στην περίπτωση της πολλαπλής. Δεν υπάρχει ωστόσο ένας συντελεστής εξάρτησης αλλά τόσοι όσες οι ανεξάρτητες μεταβλητές. Δεδομένου ότι καθένας αντιπροσωπεύει την εξάρτηση της Υ από την αντίστοιχη μεταβλητή X_i οι b_i καλούνται συντελεστές μερικής εξάρτησης.

Συντελεστής μερικής εξάρτησης :

Οι ιδιότητες των συντελεστών μερικής εξάρτησης (b_i) είναι ίδιες με αυτές που αναφέρθηκαν για τον b_1. Υπάρχει όμως διαφορά στην ερμηνεία:

Ο b_i εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής, όταν η αντίστοιχη ανεξάρτητη (X_i) μεταβληθεί κατά μια μονάδα *και όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές*. Αυτό φαίνεται από τις αντίστοιχες εξισώσεις.

Ας θεωρήσουμε το μοντέλο:

$$E(\hat{Y}_i \mid X_{1i}, X_{2i}, ..., X_{pi}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi} \qquad (7)$$

Αν υποθέσουμε ότι η X_{1i} αυξάνεται κατά μία μονάδα ενώ όλες οι υπόλοιπες παραμένουν σταθερές, η 7 γράφεται :

$$E(\hat{Y}_i \mid (X_{1i}+1), X_{2i}, ..., X_{pi}) = b_0 + b_1(X_{1i}+1) + b_2 X_{2i} + ... + b_p X_{pi} \quad (8)$$

Προκειμένου να υπολογίσουμε την μεταβολή στην μέση τιμή της Y, για μία μονάδα αύξησης της X_{1i}, αφαιρούμε τις παραπάνω σχέσεις.

$$\begin{aligned} &E(\hat{Y}_i \mid (X_{1i}+1), X_{2i}, ..., X_{pi}) - E(\hat{Y}_i \mid X_{1i}, X_{2i}, ..., X_{pi}) \\ &= b_0 + b_1(X_{1i}+1) + b_2 X_{2i} + ... + b_p X_{pi} - b_0 - b_1 X_{1i} - b_2 X_{2i} - ... - b_p X_{pi} = b_1 \end{aligned} \quad (9)$$

SPSS

Ας υποθέσουμε για παράδειγμα, ότι θέλουμε να μελετήσουμε τον τρόπο με τον οποίο μεταβάλλεται το βάρος των γυναικών *(εξαρτημένη μεταβλητή),* σε σχέση με το ύψος τους αλλά και την ηλικία τους *(ανεξάρτητες)*. Η μεταβλητή βάρος κατανέμεται κανονικά, οπότε μπορούμε να εφαρμόσουμε την μέθοδο.

Analyze → Regression → Linear → και βάζουμε την εξαρτημένη και τις ανεξάρτητες μεταβλητές μας στο κατάλληλο πλαίσιο → **Ok**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,305[a]	,093	,082	10,89

a. Predictors: (Constant), Σταθμισμένη Ηλικία, Σταθμισμένο Ύψος

σχόλιο:

$\mathbf{R^2}$ = 0.093 → Το μοντέλο επεξηγεί το 9.3% της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής.

Το $\mathbf{R^2}$ αυξήθηκε με την εισαγωγή μίας ακόμα επεξηγηματικής μεταβλητής στο μοντέλο (αυτό συμβαίνει πάντα). Το $\mathbf{R^2}$ **adjusted** παρέμεινε σταθερό, γεγονός που οφείλεται στο ότι η μεταβλητή που εισήχθη δεν επιδρά σε βαθμό στατιστικά σημαντική, όπως θα δούμε.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1923,531	2	961,765	8,104	,000[a]
	Residual	18750,643	158	118,675		
	Total	20674,174	160			

a. Predictors: (Constant), Σταθμισμένη Ηλικία, Σταθμισμένο Ύψος

b. Dependent Variable: Βάρος (πριν)

σχόλιο:

F = M. S. (regression)/ M. S. (residuals) → το μοντέλο εξηγεί σημαντικό μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής (p-value<0.0001). Ο έλεγχος γίνεται κάτω από την μηδενική υπόθεση: MS(reg)=MS(res), δηλαδή F=1.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66,557	,860		77,352	,000
	Σταθμισμένο Ύψος	,546	,136	,311	4,009	,000
	Σταθμισμένη Ηλικία	8,242E-02	,067	,095	1,226	,222

a. Dependent Variable: Βάρος (πριν)

σχόλια:

Το μοντέλο που αντιστοιχεί στον παραπάνω πίνακα είναι το:

$$E(\hat{Y}_i \mid X_i) = 62.003 + 0.546 \cdot \Upsilon\psi o\varsigma + 0.082 \cdot \Eta\lambda\iota\kappa\iota\alpha$$

Για κάθε έναν από τους συντελεστές, γίνεται ο έλεγχος one sample t - test, κάτω από την μηδενική υπόθεση H_0: *ο συντελεστής είναι ίσος με το μηδέν* (που ισοδυναμεί στη περίπτωση του b_i, με απουσία εξάρτησης).

b_1 = + 0.546 → για κάθε εκατοστό αύξηση του ύψους, αναμένεται κατά μέσο όρο αύξηση κατά 0.546 κιλά του βάρους, δεδομένου ότι η ηλικία παραμένει σταθερή. Το αποτέλεσμα αυτό είναι στατιστικά σημαντικό στο επίπεδο του 1% (p-value <0.0001).

$\mathbf{b_2}$ = + **0.082** → για κάθε έτος αύξησης της ηλικίας, αναμένεται κατά μέσο όρο αύξηση κατά 0.082 κιλά του βάρους, δεδομένου ότι το ύψος παραμένει σταθερό. Το αποτέλεσμα αυτό δεν είναι στατιστικά σημαντικό (p-value = 0.222).

$\mathbf{b_0}$ **= 66.557** → Η τιμή αυτή αντιστοιχεί στην αναμενόμενη μέση τιμή του βάρους, για τις γυναίκες ύψους 160.84 εκ. (μέση τιμή) και ηλικίας 55.25 ετών (μέση τιμή).

Ποιοτικές μεταβλητές και χρήση ψευδομεταβλητών :

Στην πολλαπλή γραμμική εξάρτηση οι ανεξάρτητες μεταβλητές μπορεί να μη είναι όλες ποσοτικές. Όταν χρησιμοποιούνται ποιοτικές μεταβλητές με περισσότερα από 2 επίπεδα απαιτείται η δημιουργία ψευδοματαβλητών (dummy variables). Για κάθε κατηγορία της ποιοτική μεταβλητή φτιάχνεται μία ψευδοματαβλητή. Κάθε ψευδοματαβλητή παίρνει την τιμή 1 όταν το άτομο ανήκει σε αυτή την κατηγορία και 0 σε οποιαδήποτε άλλη περίπτωση. Στο μοντέλο της γραμμικής εξάρτησης εισάγονται τόσες ψευδοματαβλητές όσες ο αριθμός των κατηγοριών της μεταβλητής μείον 1. Η ψευδομεταβλητή που δεν εισάγεται στο μοντέλο αποτελεί το επίπεδο αναφοράς (reference level).

SPSS

Ας υποθέσουμε για παράδειγμα, ότι θέλουμε να μελετήσουμε τον τρόπο με τον οποίο μεταβάλλεται το βάρος των γυναικών *(εξαρτημένη μεταβλητή),* σε σχέση με το ύψος (ποσοτική) τους αλλά και τη βαθμίδα εκπαίδευσης τους (ποιοτική) *(ανεξάρτητες)*. Η μεταβλητή βάρος κατανέμεται κανονικά, οπότε μπορούμε να εφαρμόσουμε την μέθοδο.

Για να μελετήσουμε την επίδραση της βαθμίδας εκπαίδευσης, θα δημιουργήσουμε τις αντίστοιχες ψευδομεταβλητές ως εξής:

Transform → Recode → Into different variables → Numeric variable: educat , **Output variable**: edu1, **Change → Old and new values → Old value**: Value: 1, **New value**: value: 1, **Add** και **All other values**, **New value**: 0, **Add→ continue → Ok**

Με τον τρόπο αυτό δημιουργήσαμε μια νέα μεταβλητή την edu1, που έχει την τιμή 1 όταν η educat=1 και 0 οπουδήποτε αλλού. Όμοια δημιουργούμε και τις μεταβλητές edu2 και edu3.

Analyze → Regression → Linear → και βάζουμε την εξαρτημένη και τις ανεξάρτητες μεταβλητές μας στο κατάλληλο πλαίσιο. ΠΡΟΣΟΧΗ: Βάζουμε μόνο τις δυο από τις τρεις ψευδομεταβλητές (π.χ. edu2, edu3) **→ Ok**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,341[a]	,116	,100	10,79

a. Predictors: (Constant), EDU3, Σταθμισμένο Ύψος, EDU2

σχόλιο:

$\mathbf{R^2}$ = 0.116 → Το μοντέλο επεξηγεί το 11.6% της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2406,825	3	802,275	6,895	,000[a]
	Residual	18267,348	157	116,353		
	Total	20674,174	160			

a. Predictors: (Constant), EDU3, Σταθμισμένο Ύψος, EDU2

b. Dependent Variable: Βάρος (πριν)

σχόλιο:

F = M. S. (regression)/ M. S. (residuals) → το μοντέλο εξηγεί σημαντικό μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής (p-value<0.0001).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	68,047	1,240		54,891	,000
	Σταθμισμένο Ύψος	,515	,133	,294	3,884	,000
	EDU2	-1,718	1,876	-,073	-,916	,361
	EDU3	-5,921	2,487	-,189	-2,381	,018

a. Dependent Variable: Βάρος (πριν)

σχόλια:

Το μοντέλο που αντιστοιχεί στον παραπάνω πίνακα είναι το:

$$E(\hat{Y}_i \mid X_{1i}, X_{2i}, X_{3i}) = 68.047 + 0.515 \cdot \Upsilon\psi o\varsigma - 1.718 \cdot Edu2 - 5.921 \cdot Edu3$$

Ο συντελεστής μερικής εξάρτησης για κάθε ψευδομεταβλητή, εκφράζει τη μέση διαφορά στην εξαρτημένη μεταβλητή για τα άτομα της κατηγορίας στην οποία αναφέρεται, από τα άτομα που ανήκουν στην κατηγορία αναφοράς. Στο συγκεκριμένο παράδειγμα κατηγορία αναφοράς αποτελεί η edu1, δηλαδή οι γυναίκες που έχουν συμπληρώσει την πρωτοβάθμια βαθμίδα εκπαίδευσης.

Για τις γυναίκες που έχουν τελειώσει τη πρωτοβάθμια εκπαίδευση (edu1) το αντίστοιχο μοντέλο είναι:

$$E(\hat{Y}_i \mid X_{1i}, X_{2i}, X_{3i})_1 = 68.047 + 0.515 \cdot \Upsilon\psi o\varsigma$$

Για τις γυναίκες που έχουν τελειώσει τη δευτεροβάθμια εκπαίδευση (edu2) το αντίστοιχο μοντέλο είναι:

$$E(\hat{Y}_i \mid X_{1i}, X_{2i}, X_{3i})_2 = 68.047 + 0.515 \cdot \Upsilon\psi o\varsigma - 1.718$$

Αφαιρώντας τις παραπάνω εξισώσεις έχουμε:

$$E(\hat{Y}_i \mid X_i)_2 - E(\hat{Y}_i \mid X_i)_1 = 68.047 + 0.515 \cdot \Upsilon\psi o\varsigma - 1.718 - 68.047 - 0.515 \cdot \Upsilon\psi o\varsigma = -1.718$$

b_3 = - 1.718 → Οι γυναίκες που έχουν τελειώσει τη δευτεροβάθμια εκπαίδευση αναμένουμε να έχουν κατά μέσο όρο μικρότερο βάρος κατά 1.7 κιλά από τις γυναίκες που έχουν τελειώσει τη πρωτοβάθμια εκπαίδευση, για σταθερό ύψος. Το αποτέλεσμα αυτό δεν είναι στατιστικά σημαντικό (p-value = 0.361).

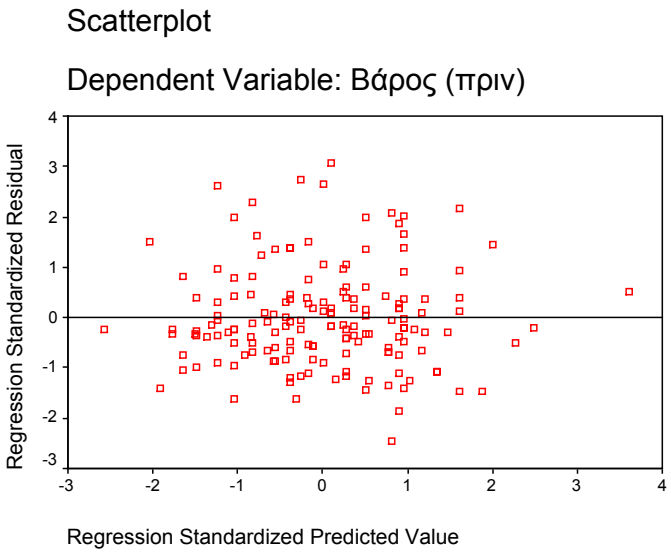
$\mathbf{b_4}$ = - 5.921 → Οι γυναίκες που έχουν τελειώσει τη τριτοβάθμια εκπαίδευση αναμένουμε να έχουν κατά μέσο όρο μικρότερο βάρος κατά 5.9 κιλά από τις γυναίκες που έχουν τελειώσει τη πρωτοβάθμια εκπαίδευση, για σταθερό ύψος. Το αποτέλεσμα αυτό είναι στατιστικά σημαντικό (p-value = 0.018).

$\mathbf{b_0}$ **= 66.557** → Οι γυναίκες που έχουν ύψος ίσο με 160.84 εκ. και έχουν τελειώσει τη πρωτοβάθμια εκπαίδευση αναμένεται να έχουν κατά μέσο όρο βάρος ίσο με 66.557

Διαγνωστικά διαγράμματα καταλοίπων :

Όπως αναφέραμε τα κατάλοιπα πρέπει να κατανέμονται κανονικά με μέση τιμή 0 και σταθερή διασπορά σ^2. Το παρακάτω διάγραμμα μπορεί να μας πληροφορήσει για το αν οι προϋποθέσεις πληρούνται ή όχι. Επίσης, έχουμε ήδη δει και τη δοκιμασία των Kolmogorov-Smirnov όπως και το ιστόγραμμα.

Analyze → Regression → Linear → Plots → X: ZPRED**, Y:**ZRESID **→ Continue → Save → Residuals**: Standardized **→ Continue → Ok.**



Λίγα λόγια για την επιλογή μοντέλου

Πολλές φορές, προκειμένου να εξηγήσουμε την μεταβλητότητα ενός μεγέθους, έχουμε στην διάθεσή μας δεδομένα για διάφορες μεταβλητές. Θα πρέπει να επιλέξουμε ποιες από αυτές θα εισάγουμε στο μοντέλο πολλαπλής παλινδρόμησης. Η επιλογή μοντέλων είναι μεγάλο κεφάλαιο της στατιστικής, για το οποίο θα αναφερθούν μόνο τα βασικά σημεία..

Το στατιστικό κριτήριο στο οποίο βασιζόμαστε, προκειμένου να αποφασίσουμε αν μία ανεξάρτητη μεταβλητή θα εισαχθεί ή όχι στο μοντέλο, είναι το *αν αυτή συνεισφέρει σε βαθμό στατιστικά σημαντικό στην επεξήγηση της μεταβλητότητας της εξαρτημένης μεταβλητής (σκοπός του μοντέλου).* Στατιστικά, αυτό ελέγχεται από την τιμή του p-value του t-test, για τον αντίστοιχο συντελεστή μερικής εξάρτησης. Νωρίτερα είδαμε:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66,557	,860		77,352	,000
	Σταθμισμένο Ύψος	,546	,136	,311	4,009	,000
	Σταθμισμένη Ηλικία	8,242E-02	,067	,095	1,226	,222

a. Dependent Variable: Βάρος (πριν)

Παρατηρούμε ότι το ύψος συνεισφέρει σε βαθμό στατιστικά σημαντικό στην εξήγηση της μεταβλητότητας τους βάρους (p – value < 0.0001) , άρα θα πρέπει να «παραμείνει» στο μοντέλο. Αντίθετα η ηλικία, δεν συνεισφέρει σε βαθμό στατιστικά σημαντικό στην εξήγηση της μεταβλητότητας τους βάρους (p – value =0.222) , άρα δεν θα εισαχθεί τελικά στο μοντέλο.

Η διαδικασία επιλογής των επεξηγηματικών μεταβλητών του μοντέλου είναι προτιμότερο να γίνεται από τον χρήστη του Spss , εισάγοντας μία μία μεταβλητή και ελέγχοντας την επίδραση της στην εξαρτημένη μεταβλητή. Ενδέχεται ωστόσο, αυτό να είναι μία χρονοβόρα διαδικασία όταν ο αριθμός των υποψήφιων μεταβλητών είναι πάρα πολύ μεγάλος. Έτσι το Spss, όπως όλα τα στατιστικά πακέτα, διαθέτει εντολές αυτόματης επιλογής μεταβλητών.

SPSS

Analyze → Regression → Linear → και βάζουμε την εξαρτημένη και την ανεξάρτητη μεταβλητή μας στο κατάλληλο πλαίσιο → **Method →** enter ή stepwise ή backword ή forword →**Ok**

- Μέθοδος enter : ο χρήστης εισάγει μόνος του τις μεταβλητές που επιθυμεί.

- Μέθοδος Forward : το Spss επιλέγει ποιες μεταβλητές θα μπουν στο μοντέλο, με κριτήριο τα αντίστοιχα b να είναι στατιστικά σημαντικά τουλάχιστον στο επίπεδο του 20% (p – value ≤0.2), ξεκινώντας από αυτή που έχει το μικρότερο p – value.

- Μέθοδος Backward : αντίθετα, το Spss εισάγει όλες τις μεταβλητές στο μοντέλο, και αφαιρεί μία μία τις στατιστικά μη σημαντικές στο επίπεδο του 20% (p – value >0.2), ξεκινώντας από αυτή που έχει το μεγαλύτερο p – value.

- Μέθοδος Stepwise : είναι συνδυασμός σε βήματα, των δύο προηγούμενων. Εισάγει και εξάγει μεταβλητές στο μοντέλο προκειμένου να καταλήξει σε αυτό με την μεγαλύτερη προγνωστική αξία, δηλαδή το μικρότερο M.S.(residuals).