Αλεξάνδρειο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Θεσσαλονίκης
Τμήμα Πληροφορικής
Εργαστήριο «Θεωρία Πιθανοτήτων και Στατιστική»

ΣΥΣΧΕΤΙΣΗ και ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Περιεχόμενα

1. Συσχέτιση μεταξύ δύο ποσοτικών μεταβλητών 2
Παράδειγμα 1 2
1.1 Δεδομένα 2
1.2 Ζητήματα 2
1.3 Διάγραμμα διασποράς 2
1.4 Συντελεστής Γραμμικής Συσχέτισης 3
1.5 Ερμηνεία αποτελεσμάτων 3
1.6 Αναφορά αποτελεσμάτων 4
2. Γραμμική Παλινδρόμηση 5
Παράδειγμα 2.1 (Απλή Παλινδρόμηση) 5
2.1.1 Δεδομένα 5
2.1.2 Ζητήματα 5
2.1.3 Επιλογή και προσδιορισμός του μαθηματικού μοντέλου 5
2.1.4 Ερμηνεία αποτελεσμάτων 7
2.1.5 Αναφορά αποτελεσμάτων 7
Παράδειγμα 2.2 (Πολλαπλή Παλινδρόμηση) 8
2.2.1 Δεδομένα 8
2.2.2 Ζητήματα 8
2.2.3 Επιλογή και προσδιορισμός του μαθηματικού μοντέλου 8
2.2.4 Αναφορά αποτελεσμάτων 9

1. Συσχέτιση μεταξύ δύο ποσοτικών μεταβλητών
(Διαδικασίες: Bivariate Correlations & Scatter/Dot)

Ο προσδιορισμός της φύσης και της έντασης της συσχέτισης μεταξύ δύο ποσοτικών μεταβλητών πραγματοποιείται με την κατασκευή του διαγράμματος διασποράς (διαδικασία **Scatterplot**) και με τον υπολογισμό του συντελεστή γραμμικής συσχέτισης του **Pearson** (διαδικασία **Bivariate Correlations**).

Παράδειγμα 1

1.1 Δεδομένα

Για το παρακάτω παράδειγμα, θα χρησιμοποιηθεί το αρχείο ERG-STAT-Askisi_3_data

1.2 Ζητήματα

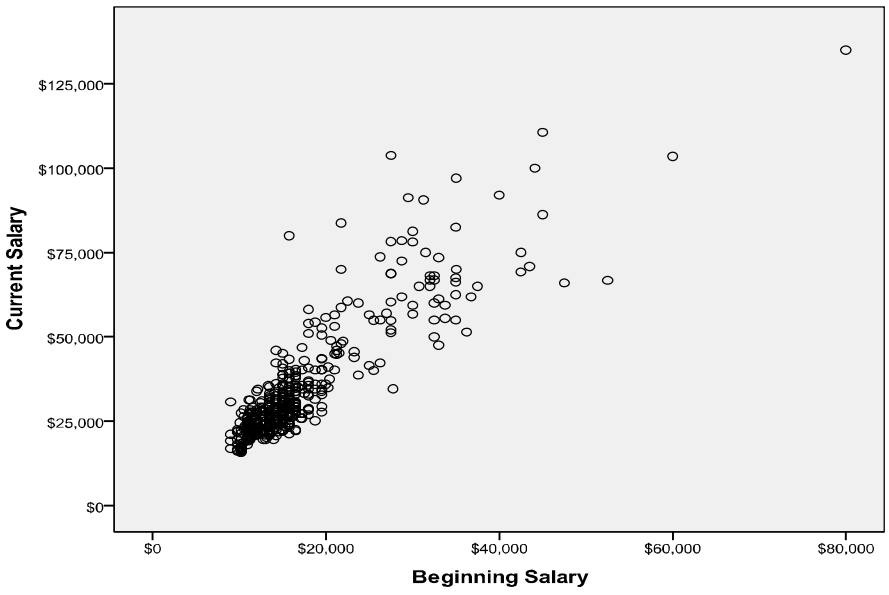
Θα θέλαμε να προσδιορίσουμε την φύση και την ένταση της συσχέτισης, η οποία ενδεχομένως να υπάρχει, μεταξύ του τρέχοντος μισθού (*current salary*) των εργαζομένων στην επιχείρηση αυτή και του αντίστοιχου αρχικού μισθού πρόσληψης τους (*beginning salary*).

1.3 Διάγραμμα διασποράς

Για τον προσδιορισμό της φύσης της συσχέτισης που ενδεχομένως υπάρχει μεταξύ των δύο χαρακτηριστικών (μεταβλητών), κατασκευάζουμε το διάγραμμα διασποράς (**Graphs - ScatterPlot**). Συγκεκριμένα, η κατασκευή αυτή πραγματοποιείται ως εξής:
Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε

Graphs – Scatter/Dot
Simple Scatter -- Define

Στο πλαίσιο διαλόγου της διαδικασίας **Simple Scatterplot**, επιλέγουμε τις δύο ποσοτικές μεταβλητές που μας ενδιαφέρουν και τις μετακινούμε στα πλαίσια **Y Axis** και **X Axis**. Όταν η μια από τις δύο μεταβλητές θεωρείται ως ανεξάρτητη μετακινείται στο πλαίσιο X Axis (στο παράδειγμα αυτό ως ανεξάρτητη θεωρείται η μεταβλητή *Beginning Salary*).



1.4 Συντελεστής Γραμμικής Συσχέτισης

Για να εκτιμήσουμε την ένταση της συσχέτισης αυτής, υπολογίζουμε τον συντελεστή γραμμικής συσχέτισης του Pearson (r). Ο **συντελεστής γραμμικής συσχέτισης του Pearson (r)** παίρνει τις τιμές: **$-1 \leq r \leq +1$**, όπου

r = -1	**r = 0**	**r = +1**
Τέλεια αρνητική γραμμική συσχέτιση	Μηδενική (δεν υπάρχει) γραμμική συσχέτιση	Τέλεια θετική γραμμική συσχέτιση

Συγκεκριμένα:
Όσο το r βρίσκεται πιο κοντά στο +1 (-1), τόσο πιο ισχυρή θετική (αρνητική) συσχέτιση υπάρχει. Όσο το r βρίσκεται πιο κοντά στο 0, τόσό πιο ασθενής συσχέτιση υπάρχει. Ενδεικτικά, μπορούμε να θεωρήσουμε ότι η συσχέτιση είναι:

Ικανοποιητική έως πολύ ισχυρή,	όταν	$0,7 < \lvert r \rvert < 1$
Μέτρια έως ικανοποιητική,	όταν	$0,5 < \lvert r \rvert < 0,7$
Ασθενής έως μέτρια,	όταν	$0 < \lvert r \rvert < 0,5$

Ο υπολογισμός του συντελεστή γραμμικής συσχέτισης πραγματοποιείται ως εξής:

Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε:
Analyze – Correlate – Bivariate
Στο πλαίσιο διαλόγου της διαδικασίας **Bivariate Correlations**, επιλέγουμε τις μεταβλητές των οποίων τη σχέση αναζητούμε και τις μετακινούμε στο πλαίσιο variable(s). Μπορούμε να μετακινήσουμε περισσότερες από δύο μεταβλητές. Στην περίπτωση αυτή οι υπολογισμοί θα γίνουν για κάθε ένα συνδυασμό τους ανά δύο.

Correlations

		Current Salary	Beginning Salary
Current Salary	Pearson Correlation	1	.880**
	Sig. (2-tailed)		.000
	N	474	474
Beginning Salary	Pearson Correlation	.880**	1
	Sig. (2-tailed)	.000	
	N	474	474

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

1.5 Ερμηνεία αποτελεσμάτων

Το διάγραμμα διασποράς υποδεικνύει την ύπαρξη γραμμικής συσχέτισης μεταξύ των δύο χαρακτηριστικών (τρέχον μισθός – αρχικός μισθός). Παρατηρούμε ότι τα σημεία τείνουν να συγκεντρώνονται γύρω από μία νοητή ευθεία. Με άλλα λόγια, παρατηρούμε ότι όταν ο αρχικός μισθός αυξάνεται, και ο τρέχον μισθός αυξάνεται επίσης, κατά μέσο όρο.
Από τον πίνακα αποτελεσμάτων που αφορά στον συντελεστή γραμμικής συσχέτισης, καταγράφουμε την τιμή του συντελεστή (r = + 0,880) και την σημαντικότητα του ελέγχου (sig.=0,000).
Ο συντελεστής (r) ερμηνεύεται σύμφωνα με όσα έχουν αναφερθεί στην προηγούμενη ενότητα.
Ο έλεγχος υποθέσεων που πραγματοποιείται αφορά στην πραγματική τιμή του συντελεστή γραμμικής συσχέτισης ρ (στον πληθυσμό). Συγκεκριμένα, ο έλεγχος υποθέσεων περιγράφεται ως εξής:

$H_0 : \rho = 0$ $H_a : \rho \neq 0$
Εφαρμογή του ελέγχου
sig. = 0,000 < 0,05, συνεπώς απορρίπτουμε τη μηδενική υπόθεση.
Δηλαδή, ο συντελεστής ρ θεωρείται ότι είναι στατιστικά σημαντικά διάφορος του μηδέν, και συνεπώς υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ του αρχικού και του τρέχοντος μισθού.

Παρατηρήσεις
α) Η διαφορά μεταξύ του ρ και του r είναι ότι το πρώτο αναφέρεται στον πραγματικό συντελεστή συσχέτισης (δηλαδή εκείνον που θα προέκυπτε εάν είχαμε δεδομένα από το σύνολο του "πληθυσμού" των εργαζόμενων, και την τιμή του οποίου δεν γνωρίζουμε και θέλουμε να εκτιμήσουμε), ενώ το r αναφέρεται στον συντελεστή συσχέτισης που υπολογίζεται από τα δεδομένα του συγκεκριμένου δείγματος εργαζομένων. Συνεπώς, πρώτα ελέγχουμε εάν ο συντελεστής ρ μπορεί να θεωρηθεί ότι είναι στατιστικά σημαντικά διάφορος από το μηδέν (0), και μόνο στην περίπτωση αυτή προχωρούμε στην καταγραφή της έντασης της συσχέτισης αυτής με την ερμηνεία της τιμής του r.
β) Στην περίπτωση κατά την οποία η τιμή του συντελεστή **ΓΡΑΜΜΙΚΗΣ** συσχέτισης (r) βρίσκεται κοντά στο 0 κατά απόλυτη τιμή, μπορούμε να συμπεράνουμε ότι οι δύο μεταβλητές δεν συσχετίζονται μεταξύ τους **ΓΡΑΜΜΙΚΑ** (δηλαδή η σχέση τους δεν μπορεί να περιγραφεί ικανοποιητικά από μία ευθεία). Αυτό όμως δεν αποκλείει ότι μπορεί να συσχετίζονται με κάποιον άλλο τρόπο, **ΜΗ ΓΡΑΜΜΙΚΟ**. Για το λόγο αυτό, είναι απαραίτητο να κατασκευάζουμε πρώτα το διάγραμμα διασποράς, το οποίο μπορεί να υποδείξει την φύση της συσχέτισης που μπορεί να υφίσταται μεταξύ δύο μεταβλητών.

1.6 Αναφορά αποτελεσμάτων

Οι παραπάνω ερμηνείες μπορούν να καταγραφούν σε μία αναφορά αποτελεσμάτων ως εξής:
Το διάγραμμα διασποράς υποδεικνύει την ύπαρξη γραμμικής συσχέτισης μεταξύ των δύο χαρακτηριστικών (τρέχον μισθός – αρχικός μισθός). Συγκεκριμένα, παρατηρούμε ότι όταν ο αρχικός μισθός αυξάνεται, και ο τρέχον μισθός αυξάνεται επίσης, κατά μέσο όρο.
Σύμφωνα με τον συντελεστή γραμμικής συσχέτισης (r = +0,880, sig.<0,000), θεωρούμε ότι υπάρχει αρκετά ισχυρή, θετική, γραμμική συσχέτιση μεταξύ του αρχικού και του τρέχοντος μισθού.

2. Γραμμική Παλινδρόμηση
(Διαδικασία: Linear Regression)

Η Γραμμική Παλινδρόμηση αποτελεί μία στατιστική μέθοδο η οποία αποσκοπεί στον προσδιορισμό ενός μαθηματικού μοντέλου για την περιγραφή / ερμηνεία / πρόβλεψη των τιμών ενός χαρακτηριστικού (μεταβλητής) σε σχέση με τις τιμές ενός πλήθους άλλων χαρακτηριστικών (μεταβλητών).

Παράδειγμα 2.1 (Απλή Παλινδρόμηση)

2.1.1 Δεδομένα

Για το παρακάτω παράδειγμα, θα χρησιμοποιηθεί το αρχείο ERG-STAT-Askisi_3_data

2.1.2 Ζητήματα

Θα θέλαμε να προσδιορίσουμε ένα μαθηματικό μοντέλο το οποίο να περιγράφει τον τρέχοντα μισθό των εργαζόμενων (*current salary*) σε σχέση με τον αντίστοιχο αρχικό μισθό πρόσληψης τους (*beginning salary*). Το ζήτημα αυτό βασίζεται στην λογική υπόθεση ότι ο τρέχων μισθός ενός εργαζόμενου μπορεί να «εξαρτάται» από το μισθό πρόσληψης (αρχικό μισθό).

2.1.3 Επιλογή και προσδιορισμός του μαθηματικού μοντέλου

Όπως έχουμε διαπιστώσει από το προηγούμενο παράδειγμα, ο τρέχων και ο αρχικός μισθός συσχετίζονται αρκετά ισχυρά και γραμμικά μεταξύ τους. Συνεπώς, το μαθηματικό μοντέλο, το οποίο μπορούμε να επιλέξουμε και να προσαρμόσουμε στα δεδομένα μας είναι το γραμμικό.

Το γραμμικό μοντέλο:

$$\mathbf{Y = \alpha + \beta X + \varepsilon}$$

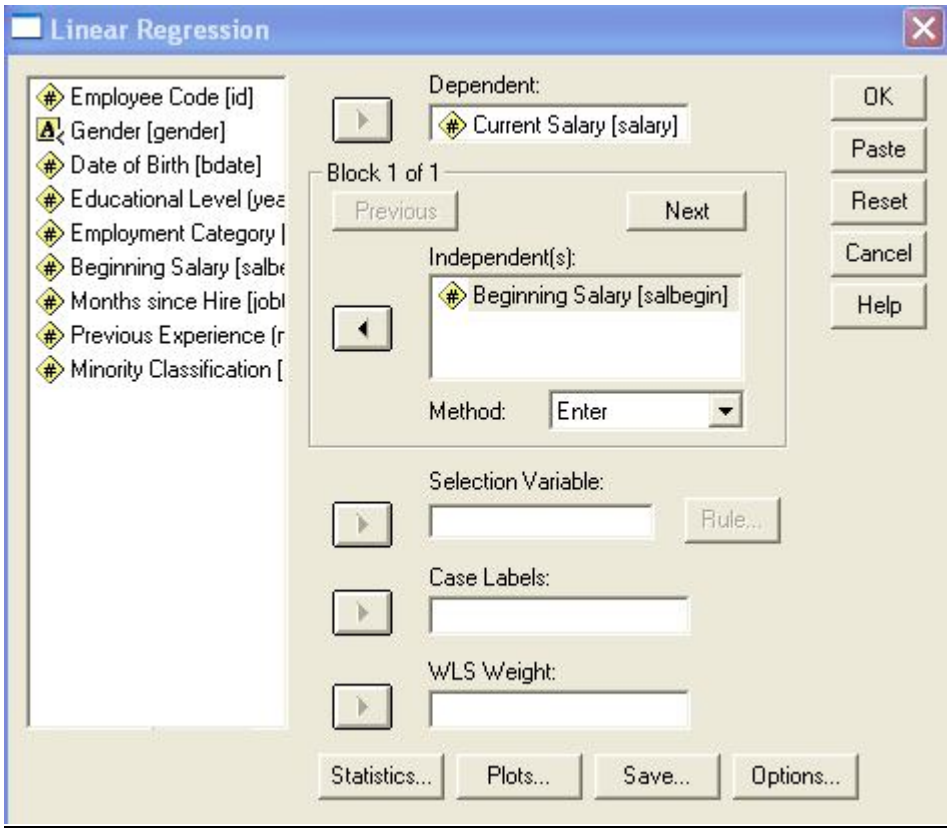
εξαρτημένη μεταβλητή ανεξάρτητη μεταβλητή σφάλμα

όπου, α β και ε πραγματικοί αριθμοί

Ο προσδιορισμός και η αξιολόγηση του εν λόγω μοντέλου (Γραμμή Παλινδρόμησης) πραγματοποιείται ως εξής:
Από τη βασική ράβδο προτιμήσεων του λογισμικού επιλέγουμε:

Analyze – Regression – Linear

Μεταφέρουμε τη μεταβλητή την οποία μελετούμε (ή / και θέλουμε να κάνουμε πρόβλεψη των τιμών της) στο πλαίσιο **Dependent** (*current salary*) και την μεταβλητή, την οποία θα χρησιμοποιήσουμε για να ερμηνεύσουμε τις τιμές της πρώτης, στο πλαίσιο **Independent(s)** (*beginning salary*). **OK**.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.880[a]	.775	.774	$8,115.356

a. Predictors: (Constant), Beginning Salary

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1E+011	1	1.068E+011	1622.118	.000[a]
	Residual	3E+010	472	65858997.22		
	Total	1E+011	473			

a. Predictors: (Constant), Beginning Salary

b. Dependent Variable: Current Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1928.206	888.680		2.170	.031
	Beginning Salary	1.909	.047	.880	40.276	.000

a. Dependent Variable: Current Salary

2.1.4 Ερμηνεία αποτελεσμάτων

Πρώτος Πίνακας

Ο δείκτης R-square (R^2 = 0,775) εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής το οποίο ερμηνεύεται από τη διακύμανση των τιμών της ανεξάρτητης μεταβλητής. Δηλαδή στο παράδειγμα, το 77,5% της διακύμανσης των μισθών των εργαζομένων ερμηνεύεται από τη διακύμανση των αρχικών μισθών τους.

Ο συντελεστής αυτός ονομάζεται συντελεστής προσδιορισμού και υποδεικνύει την ποιότητα προσαρμογής της γραμμής παλινδρόμησης στα δεδομένα.

Δεύτερος Πίνακας

Στον πίνακα των αποτελεσμάτων της Ανάλυσης Διακύμανσης (ANOVA), παρατηρείται ότι το επίπεδο σημαντικότητας είναι: sig.=0,000 < 0,05. Συνεπώς, η γραμμή παλινδρόμησης που έχει εκτιμηθεί θεωρείται ότι είναι στατιστικά σημαντική (F=1622,118 , sig. = 0,000).

Τρίτος Πίνακας

Το μαθηματικό μοντέλο το οποίο προκύπτει (γραμμή παλινδρόμησης), σύμφωνα με τον τρίτο πίνακα, είναι το εξής:

(Current Salary) = 1928,206 + 1,909 (Beginning Salary)

Έλεγχος του συντελεστή παλινδρόμησης β:

$H_0 : \beta = 0$ $H_a : \beta \neq 0$

Εφαρμογή του t-test

sig = 0,000 < 0,05, συνεπώς απορρίπτουμε τη μηδενική υπόθεση

Ο συντελεστής παλινδρόμησης β είναι στατιστικά σημαντικά διάφορος του μηδέν (t = 40,276 , sig < 0,001), συνεπώς ο αρχικός μισθός ερμηνεύει στατιστικά σημαντικά το μισθό των εργαζομένων.

Ερμηνεία του συντελεστή παλινδρόμησης β

Δεδομένου ότι ο συντελεστής αυτός είναι στατιστικά σημαντικός, η τιμή του ερμηνεύεται ως εξής: όταν ο αρχικός μισθός είναι αυξημένος κατά μία μονάδα (1 δολάριο), τότε ο μισθός αναμένεται να είναι αυξημένος κατά 1,909 μονάδες (1,909 δολάρια, δηλαδή σχεδόν κατά δύο δολάρια).

2.1.5 Αναφορά αποτελεσμάτων

Οι παραπάνω ερμηνείες μπορούν να καταγραφούν σε μία αναφορά αποτελεσμάτων ως εξής:
Ο τρέχων μισθός των εργαζομένων σε σχέση με τον αρχικό μισθό τους περιγράφεται σύμφωνα με την εξής γραμμή παλινδρόμησης: (Current Salary) = 1928,206 + 1,909 (Beginning Salary).
Η γραμμή παλινδρόμησης που έχει εκτιμηθεί θεωρείται ότι είναι στατιστικά σημαντική (F=1622,118 , sig.=0,000). Συγκεκριμένα, ο αρχικός μισθός ερμηνεύει στατιστικά σημαντικά το μισθό των εργαζομένων (t = 40,276 , sig < 0,001). Το 77,5% της διακύμανσης του τρέχοντος μισθού των εργαζομένων ερμηνεύεται από τη διακύμανση των αρχικών μισθών τους.
Επίσης, το εκτιμώμενο μοντέλο υποδεικνύει ότι, όταν ο αρχικός μισθός είναι αυξημένος κατά μία μονάδα (1 δολάριο), τότε ο τρέχων μισθός αναμένεται να είναι, κατά μέσο όρο, αυξημένος κατά 1,909 μονάδες (1,909 δολάρια, δηλαδή σχεδόν κατά δύο δολάρια).

Παράδειγμα 2.2 (Πολλαπλή Παλινδρόμηση)

2.2.1 Δεδομένα

Το παράδειγμα αυτό αποτελεί συνέχεια του προηγούμενου παραδείγματος (Παράδειγμα 2.1) και βασίζεται στο ίδιο αρχείο (ERG-STAT-Askisi_3_data)

2.2.2 Ζητήματα

Θα θέλαμε να προσδιορίσουμε ένα μαθηματικό μοντέλο το οποίο να περιγράφει τον τρέχοντα μισθό των εργαζόμενων (*current salary*) σε σχέση με τον αντίστοιχο αρχικό μισθό πρόσληψης τους (*beginning salary*) και την προϋπηρεσία τους (*previous experience*). Το ζήτημα αυτό βασίζεται στην λογική υπόθεση ότι ο τρέχων μισθός ενός εργαζόμενου μπορεί να «εξαρτάται» από τον αρχικό μισθό πρόσληψης και από την προϋπηρεσία τους.

2.2.3 Επιλογή και προσδιορισμός του μαθηματικού μοντέλου

Το γραμμικό μοντέλο:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

εξαρτημένη μεταβλητή ανεξάρτητες μεταβλητές σφάλμα

όπου, α β και ε πραγματικοί αριθμοί

Εκτιμούμε το μαθηματικό μοντέλο αυτό όπως το πραγματοποιήσαμε στην προηγούμενο παράδειγμα (Παράδειγμα 2.1). Η μόνη διαφορά είναι ότι στο πλαίσιο **Independent(s)** εισάγουμε τις μεταβλητές *beginning salary* και *previous experience*.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.891[a]	.793	.793	$7,776.652

a. Predictors: (Constant), Previous Experience (months), Beginning Salary

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.094E11	2	5.472E10	904.752	.000[a]
	Residual	2.848E10	471	6.048E7		
	Total	1.379E11	473			

a. Predictors: (Constant), Previous Experience (months), Beginning Salary

b. Dependent Variable: Current Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3850.718	900.633		4.276	.000
	Beginning Salary	1.923	.045	.886	42.283	.000
	Previous Experience (months)	-22.445	3.422	-.137	-6.558	.000

a. Dependent Variable: Current Salary

2.2.4 Αναφορά αποτελεσμάτων

Οι ερμηνείες των αποτελεσμάτων προκύπτουν σύμφωνα με όσα έχουν αναφερθεί στο προηγούμενο παράδειγμα (Παράδειγμα 2.1), ενώ η καταγραφή τους σε μία αναφορά αποτελεσμάτων μπορεί να πραγματοποιηθεί ως εξής:
Ο τρέχων μισθός των εργαζομένων σε σχέση με τον αρχικό μισθό τους και την προϋπηρεσία τους περιγράφεται σύμφωνα με την εξής γραμμή παλινδρόμησης:
(Current Salary) = 3850,718 + 1,923 (Beginning Salary) – 22,445 (Previous Experience).
Η γραμμή παλινδρόμησης που έχει εκτιμηθεί θεωρείται ότι είναι στατιστικά σημαντική (F=904,752, sig.=0,000). Συγκεκριμένα, ο αρχικός μισθός ερμηνεύει στατιστικά σημαντικά το μισθό των εργαζομένων (t = 42,283 , sig < 0,001), όπως επίσης και η προϋπηρεσία τους (t = -6,558, sig < 0,001). Το 79,3% της διακύμανσης του τρέχοντος μισθού των εργαζομένων ερμηνεύεται από τη διακύμανση του αρχικού μισθού τους και της αντίστοιχης προϋπηρεσίας τους.
Επίσης, το εκτιμώμενο μοντέλο υποδεικνύει ότι, όταν ο αρχικός μισθός είναι αυξημένος κατά μία μονάδα (1 δολάριο) και η αντίστοιχη προϋπηρεσία παραμένει σταθερή, τότε ο τρέχων μισθός αναμένεται να είναι, κατά μέσο όρο, αυξημένος κατά 1,923 μονάδες (1,923 δολάρια). Αντίστοιχα, , όταν ο αρχικός μισθός παραμένει σταθερός και η αντίστοιχη προϋπηρεσία είναι αυξημένη κατά μία μονάδα (1 μήνα), τότε ο τρέχων μισθός αναμένεται να είναι, κατά μέσο όρο, μειωμένος κατά 22,445 μονάδες (22,445 δολάρια).

Παρατήρηση
Η εισαγωγή ενός νέου ανεξάρτητου χαρακτηριστικού (μεταβλητής) στο μοντέλο της γραμμικής παλινδρόμησης βελτίωσε την ερμηνευτική ικανότητα του (R-square) κατά (μόνο) 1,8 ποσοστιαίες μονάδες!