ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Δρ. Βασίλης Π. Αγγελίδης
Τμήμα Μηχανικών Παραγωγής & Διοίκησης
Δημοκρίτειο Πανεπιστήμιο Θράκης

Πολλαπλή Παλινδρόμηση

Πολλαπλή Παλινδρόμηση

Όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές και θέλουμε να εξετάσουμε την επίδραση τους σε μία εξαρτημένη μεταβλητή χρησιμοποιούμε την **πολλαπλή γραμμική παλινδρόμηση**.

Να τονίσουμε ότι όταν χρησιμοποιούμε το όρο "γραμμική", εννοούμε "γραμμική" ως προς τις παραμέτρους του μοντέλου **(β_i).** Άρα η συνάρτηση της ευθείας ελαχίστων τετραγώνων για την περίπτωση της πολλαπλής γραμμικής παλινδρόμησης θα είναι της μορφής:

$$E\left(\hat{Y}_i \middle/ X_{1i}, X_{2i}, X_{3i}, \ldots\ldots X_{pi}\right) = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + \ldots + b_p X_{pi}$$

Όσα αναφέρθηκαν στην προηγούμενη παράγραφο όσον αφορά τις προϋποθέσεις εφαρμογής της παλινδρόμησης, ισχύουν και στην περίπτωση της πολλαπλής.

Συντελεστές μερικής εξάρτησης

Δεν υπάρχει ωστόσο ένας συντελεστής εξάρτησης αλλά τόσοι όσες οι ανεξάρτητες μεταβλητές. Δεδομένου ότι καθένας αντιπροσωπεύει την εξάρτηση της Υ από την αντίστοιχη μεταβλητή **Xi** οι **bi** καλούνται συντελεστές μερικής εξάρτησης

Οι ιδιότητες των συντελεστών μερικής εξάρτησης (**bi**) είναι ίδιες με αυτές που αναφέρθηκαν για τον **b1**. Υπάρχει όμως διαφορά στην ερμηνεία:

- Ο **bi** εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής, όταν η αντίστοιχη ανεξάρτητη (**Xi**) μεταβληθεί κατά μια μονάδα *και* ***όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές***.

Ένα επιπλέον πρόβλημα

Οι υποθέσεις που πρέπει να ικανοποιούνται είναι οι ίδιες με την απλή γραμμική παλινδρόμηση. Μία απαραίτητη προϋπόθεση η οποία είναι απαραίτητη γενικά σε όλα τα μοντέλα με περισσότερες εκ της μίας ανεξάρτητων μεταβλητών είναι η έλλειψη συγγραμμικότητας. Η συγγραμικότητα είναι ένα σοβαρό πρόβλημα για την πολλαπλή γραμμική παλινδρόμηση. Όταν μία ανεξάρτητη μεταβλητή συσχετίζεται με μία άλλη ανεξάρτητη, δηλαδή **μέσω της μίας μπορούμε να εκτιμήσουμε τις τιμές της άλλης** τότε μιλάμε για πρόβλημα συγγραμμικότητας. Επομένως η ύπαρξη και των δύο μεταβλητών στο μοντέλο δεν είναι σωστή.

Για παράδειγμα την περίπτωση στην οποία έχουμε δύο ανεξάρτητες μεταβλητές, το βάρος και το ύψος και ενδιαφερόμαστε να δούμε πως επιδρούν πάνω σε μία εξαρτημένη μεταβλητή. Είναι προφανές ότι υπάρχει σχέση μεταξύ βάρους και ύψους. Επομένως δε χρειάζεται να γνωρίζουμε και τις δύο μεταβλητές, αφού η γνώση της μίας είναι αρκετή (μέσω της μίας μπορούμε να εκτιμήσουμε τις τιμές της άλλης).

Έλεγχοι συγγραμμικότητας

Η τοποθέτηση "**άχρηστων**" μεταβλητών στο μοντέλο μπορεί φαινομενικά να είναι καλή αλλά ουσιαστικά οδηγεί στο λεγόμενο πρόβλημα της υπερπροσαρμογής του μοντέλου.

Με το να κρατήσουμε δηλαδή και τις δύο μεταβλητές που αναφέραμε σε ένα μοντέλο, φαινομενικά το βελτιώνουμε αλλά ουσιαστικά το χειροτερεύουμε.

Οπότε ή αφαιρούμε μία εκ των δύο ή χρησιμοποιούμε άλλες τεχνικές, π.χ. κεντροποίηση των τιμών των μεταβλητών, πριν την πολλαπλή γραμμική παλινδρόμηση ή άλλες τεχνικές αντί της παλινδρόμησης.

Ένα μέτρο διάγνωσης που προσφέρεται από το SPSS είναι το **VIF**, το οποίο θα το δούμε παρακάτω.

Άλλος τρόπος είναι η ύπαρξη υψηλών τιμών του **συντελεστή γραμμικής συσχέτισης**, το **Added Variable Plot** και η **παλινδρόμηση** ανάμεσα σε ζεύγη ανεξάρτητων μεταβλητών, για τις οποίες υποψιαζόμαστε συγγραμμικότητα

Δείκτες ελέγχου συγγραμμικότητας

Στις τελευταίες δύο στήλες του πίνακα των συντελεστών της παλινδρόμησης εμφανίζονται δύο δείκτες για τον έλεγχο της συγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών. Ο δείκτης **Tolerance** που παίρνει τιμές στο διάστημα [**0-1**] και για μικρές τιμές (κοντά στο 0) η μεταβλητή είναι σχεδόν σε γραμμικό συνδυασμό με τις άλλες ανεξάρτητες μεταβλητές. Τιμές **Tolerance** **μικρότερες** του **0.5** αποτελούν ένδειξη του προβλήματος. Πιο συγκεκριμένα ισχύει ότι (1-Tolerance)% μιας μεταβλητής είναι το ποσοστό της μεταβλητικότητας της μεταβλητής το οποίο μπορούν να εξηγήσουν οι υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου.

Ο δείκτης **VIF** (Variation Inflation Factor) είναι μέτρο διάγνωσης συγγραμμικότητας. Ο δείκτης **VIF** μεγαλώνει όταν ο δείκτης **Tolerance** μικραίνει. Συνήθως, ένα πρώτο φίλτρο για τον δείκτη **VIF** αποτελεί η τιμή 5, ενώ ένα δεύτερο πιο ελαστικό φίλτρο είναι η τιμή 10. Ανεξάρτητες μεταβλητές με δείκτη **VIF** μεγαλύτερο του 10 συνιστάται να αποβάλλονται από το μοντέλο.

Στον πίνακα **Collinearity Diagnostics** έχουμε το δείκτη **Eigenvalue** (ιδιοτιμή), του οποίου οι τιμές οι οποίες πλησιάζουν προς το 0 δείχνουν μεγάλη **διασυσχέτιση (intercorrelate)**

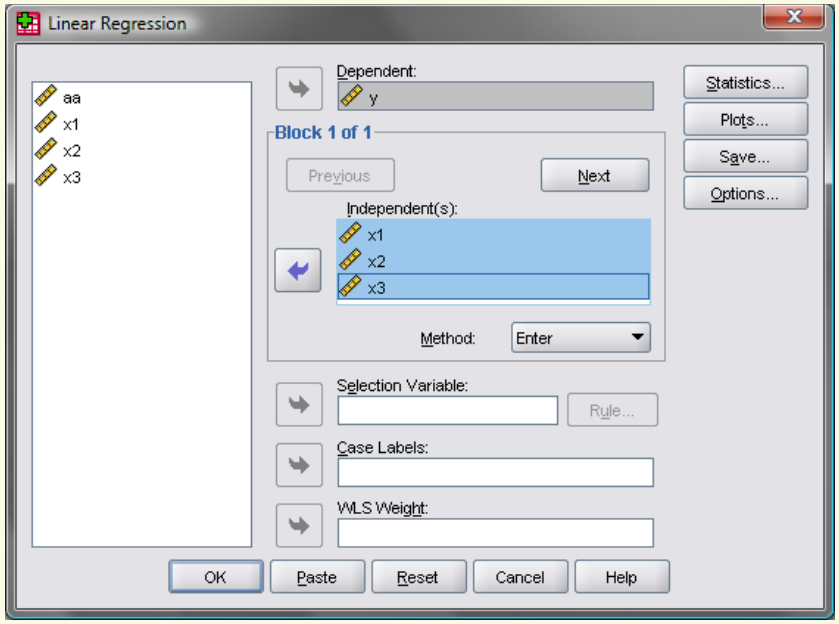
Επίσης η στήλη **Condition Index** του ίδιου πίνακα αποτελεί ένα ακόμα διαγνωστικό του προβλήματος. Τιμές **μεγαλύτερες** του **15** φανερώνουν πιθανό πρόβλημα συγγραμμικότητας και τιμές **άνω** του **30** σοβαρό πρόβλημα συγγραμμικότητας.

Παράδειγμα

Η επιτυχία μιας ομάδας η οποία συνεργάζεται για την ολοκλήρωση ενός project σίγουρα βασίζεται σε πολλούς παράγοντες. Ένας project manager για την επίλυση αυτού του προβλήματος εξετάζει την αποτελεσματικότητα τριών τεστ ικανότητας και συμπεριφοράς, στα οποία υποβάλλονται υποψήφιοι πριν την συγκρότηση της ομάδος. Σκοπός του είναι να διαπιστώσει: **α)** αν υπάρχει κάποια σχέση η οποία συνδέει τα σκορ στα τεστ αυτά με την απόδοση των υποψηφίων στην εργασία τους όταν συγκροτήσουν μια ομάδα, **β)** ποιά από τα τεστ υπεισέρχονται στη σχέση αυτή, **γ)** ποιά είναι η μορφή της σχέσης. Η απόδοση των υποψηφίων στην εργασία τους, αφού προσληφθούν, μετράται με ένα τεστ εργασιακής απόδοσης. Ένα τυχαίο δείγμα 30 υποψηφίων υποβάλλονται στα τρία τεστ και καταγράφονται τα σκορ τους (με κωδικές ονομασίες X1, X2, X3). Κατόπιν επιλέγονται για την συγκρότηση των ομάδων και μετά από ένα καθορισμένο διάστημα (ίδιο για όλους) υποβάλλονται στο τεστ εργασιακής απόδοσης. Το σκορ τους σε αυτό (κωδικός Y) καταγράφεται επίσης. Από εσάς ζητείται να βοηθήσετε το project manager, με χρήση κατάλληλης στατιστικής μεθοδολογίας, να φθάσει στα κατάλληλα συμπεράσματα (οι τιμές των μεταβλητών βρίσκονται στο αρχείο test7.sav).

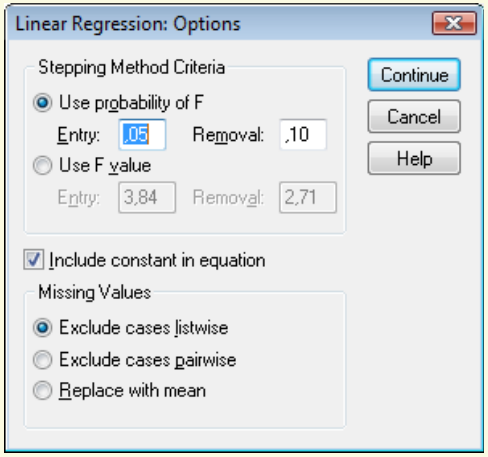
Διαδικασία πολλαπλής παλινδρόμησης

- Από το μενού **Analyze** επιλέγουμε **Regression** και στη συνέχεια **Linear**
- Επιλέγουμε την **εξαρτημένη** μεταβλητή και τη μεταφέρουμε στο παράθυρο **dependent**.
- Επιλέγουμε τις **ανεξάρτητες** μεταβλητές και τις μεταφέρουμε στο παράθυρο **independent**, και στην συνέχεια πατάμε κουμπί **Options**



Στη φόρμα αυτή έχουμε ουσιαστικά τα **κριτήρια** εισόδου-εξόδου των μεταβλητών στο μοντέλο.

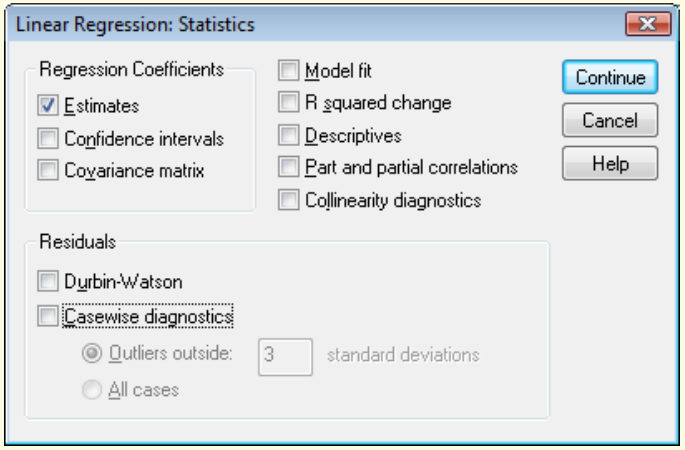
Η ένδειξη **Include constant in equation** τσεκάρεται για να πάρουμε το σταθερό όρο του μοντέλου της παλινδρόμηση



Κριτήρια εισόδου – εξόδου μεταβλητών στο μοντέλο

- Ένα κριτήριο μπορεί να είναι το **επίπεδο σημαντικότητας** (significance level) της τιμής F. Με βάση το κριτήριο αυτό, η μεταβλητή μπαίνει στο μοντέλο αν το επίπεδο σημαντικότητας (significance level) για κάθε F τιμή είναι μικρότερο από την τιμή που δώσαμε στο παράθυρο Entry. Αντίθετα αφαιρείται αν το επίπεδο σημαντικότητας (significance level) για κάθε F τιμή είναι μεγαλύτερο από την τιμή που δώσαμε στο παράθυρο Removal. Συνήθως, για να βάλουμε περισσότερες μεταβλητές στο μοντέλο μεγαλώνουμε την τιμή Entry, ενώ για να αφαιρέσουμε περισσότερες μεταβλητές μικραίνουμε την τιμή Removal.
- Ένα άλλο κριτήριο μπορεί να είναι **η τιμή F** (Use F Value). Με βάση το κριτήριο αυτό η μεταβλητή μπαίνει στο μοντέλο αν κάθε F τιμή είναι μεγαλύτερη από την τιμή που δώσαμε στο παράθυρο Entry. Αντίθετα αφαιρείται αν κάθε F τιμή είναι μικρότερη από την τιμή που δώσαμε στο παράθυρο Removal Συνήθως για να βάλουμε περισσότερες μεταβλητές στο μοντέλο μικραίνουμε την τιμή Entry, ενώ για να αφαιρέσουμε περισσότερες μεταβλητές μεγαλώνουμε την τιμή Removal.

Παράθυρο διαλόγου statistics

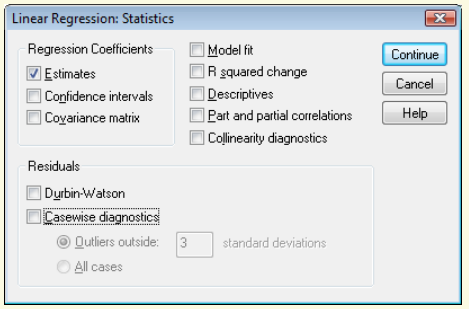


Πατώντας **statistics** μπορούμε να επιλέξουμε όλα τα στατιστικά μέτρα που επιθυμούμε

Η επιλογή **Model fit** μας δίνει τον συνοπτικό πίνακα (**Model Summary**) ο οποίος περιέχει το συντελεστή πολλαπλής συσχέτισης (R), το δείκτη προσδιορισμού (R Square), το διορθωμένο δείκτη προσδιορισμού (Adjusted R Square), το τυπικό σφάλμα της εκτίμησης (Std. Error of the Estimate) και τον πίνακα **ANOVA** στον οποίο έχουμε το άθροισμα τετραγώνων της παλινδρόμησης (regression), το άθροισμα τετραγώνων των σφαλμάτων (residual) και το συνολικό άθροισμα τετραγώνων (total). Τους βαθμούς ελευθερίας df (k, n-k-1, n-1) αντίστοιχα, τον μέσο των προηγούμενων αθροισμάτων (sum. Of square/df) Την τιμή του F κριτηρίου (Mean square Regression/Mean square Residual) και τέλος Το Sig. (περιθώριο λάθους της εκτίμησης).

Η επιλογή **Estimates** μας δίνει τον πίνακα συντελεστών ο οποίος περιέχει τον σταθερό όρο τους συντελεστές μερικής παλινδρόμησης, τις τιμές του t-test και το Sig., με βάση το οποίο δεχόμαστε ή απορρίπτουμε την μηδενική υπόθεση τη σχετική με τους συντελεστές

Παράθυρο διαλόγου statistics



Η επιλογή **part and partial correlation** μας δίνει τους μερικούς συντελεστές της παλινδρόμησης Η επιλογή **descriptives** μας εμφανίζει στα αποτελέσματα, τον πίνακα **descriptive statistics** ο οποίος περιέχει η μέση τιμή τη τυπική απόκλιση όλων των μεταβλητών και στον πίνακα **Correlation** μάς δίνονται οι **συντελεστές συσχέτισης** μεταξύ όλων των μεταβλητών (εξαρτημένης - ανεξαρτήτων) καθώς επίσης και τα **Sig.** των τεστ για όλους τους συντελεστές συσχέτισης που υπολογίστηκαν.

Η επιλογή **Confidence intervals** προσθέτει στον πίνακα **Coefficients** τα κατώτερα και ανώτερα άκρα του διαστήματος εμπιστοσύνης των συντελεστών που υπολογίσαμε.

Η επιλογή **collinearity diagnoses** προσθέτει στον πίνακα **Coefficients** τις στατιστικές συσχετίσεις, **Tolerance** (ανεκτικότητα) και **VIF** (Variance Inflation Factor) και προσθέτει τον πίνακα **Collinearity Diagnostics** που μας χρησιμεύουν για τον έλεγχο της συγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών.

Τέλος η επιλογή **Durbin Watson** προσθέτει στον πίνακα model Summary τον αντίστοιχο δείκτη για τον έλεγχο αυτοσυσχετίσεων στα κατάλοιπα.

Model Summary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,807[a]	,652	,612	10,13071	1,575

a. Predictors: (Constant), x3, x1, x2

b. Dependent Variable: y

- **R=0,807 :** Η υψηλή τιμή του συντελεστή πολλαπλής συσχέτισης του Pearson δίνει ενδείξεις γραμμικότητας του μοντέλου. Όπως έχουμε αναφέρει ο συντελεστής πολλαπλής συσχέτισης αποτελεί μέτρο της συνολικής γραμμικής σχέσης που υπάρχει μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών.
- **R^2 = 0,652** → Οι ανεξάρτητες μεταβλητές δηλαδή ο βαθμός των τριών τεστ αξιολόγησης μπορούν να ερμηνεύσουν σε ποσοστό 65,2% την συνολική μεταβλητότητα της εξαρτημένης μεταβλητής δηλαδή του βαθμού απόδοσης των εργαζομένων.
- Ο συντελεστής **Durbin Watson** ο οποίος ελέγχει την αυτοσυσχέτιση στα κατάλοιπα είναι ~**1,57** όχι τόσο καλός αλλά εφόσον δεν είναι μικρότερος του 1 ή μεγαλύτερος του 3 δεν μας ανησυχεί ιδιαίτερα.

Πίνακας Anova

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4994,476	3	1664,825	16,221	,000[a]
	Residual	2668,411	26	102,631		
	Total	7662,887	29			

a. Predictors: (Constant), x3, x1, x2 b. Dependent Variable: y

- Η τιμή **F = 16,21** (F κατανομή) είναι το πηλίκο των τιμών της στήλης Mean square.
- Η τιμή **p value**, που εμφανίζεται στην τελευταία στήλη, είναι η κρίσιμη τιμή, με την οποία αποδεχόμαστε ή απορρίπτουμε την μηδενική υπόθεση. Εδώ θα πρέπει να αναφέρουμε ότι κατά τη μηδενική υπόθεση **Ho** $(\beta_1=\beta_2=\beta_3=0)$ δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών, ενώ κατά την εναλλακτική υπόθεση H_1 υπάρχει γραμμική σχέση.
- Δηλαδή στην συγκεκριμένη περίπτωση επειδή η τιμή **p value** είναι μικρότερη του **0,05** μπορούμε να ισχυριστούμε ότι τα τρία τεστ αξιολόγησης συνδυασμένα γραμμικά με την εξίσωση της γραμμικής παλινδρόμησης που θα προκύψει από τον επόμενο πίνακα συμβάλουν σημαντικά στην ερμηνεία της μεταβλητικότητας του βαθμού απόδοσης των υπαλλήλων.

Πίνακας συντελεστών

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-28,877	19,735		-1,463	**,155**		
	x1	,328	4,460	,012	,073	**,942**	,546	1,832
	x2	3,912	1,248	,512	3,133	**,004**	,502	1,992
	x3	19,671	8,629	,367	2,280	**,031**	,516	1,938

a. Dependent Variable: y

Από τα αποτελέσματα του παραπάνω πίνακα βλέπουμε αρχικά ότι η σταθερά δεν είναι στατιστικά σημαντική για το μοντέλο και οπότε μπορούμε να την αφαιρέσουμε (η αντίστοιχη τιμή του **p – value** είναι μεγαλύτερη από το 0,05).

Όσον αφορά τις ανεξάρτητες μεταβλητές βλέπουμε ότι η πρώτη μεταβλητή, δηλαδή, το πρώτο τεστ αξιολόγησης, δεν είναι στατιστικά σημαντικό στην ερμηνεία της μεταβλητικότητας της εξαρτημένης μεταβλητής, δηλαδή, του βαθμού απόδοσης των υπαλλήλων, οπότε μπορούμε να την αφαιρέσουμε από το μοντέλο και να εκτελέσουμε ξανά την ανάλυση. Οι υπόλοιπες δύο ανεξάρτητες μεταβλητές φαίνεται να επιδρούν με στατιστικά σημαντικό τρόπο στην διαμόρφωση της εξαρτημένης μεταβλητής.

Συγγραμμικότητα

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	x1	x2	x3	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	1	3,915	1,000	,00	,00	,00	,00		
	2	,065	7,788	,03	,01	,57	,00	,546	1,832
	3	,017	15,397	,07	,02	,28	,95	,502	1,992
	4	,004	33,134	,89	,98	,15	,05	,516	1,938

a. Dependent Variable: y

Από το πίνακα των συντελεστών και ειδικά από τους δείκτες **tolerance** και **VIF** δεν εμφανίζονται ιδιαίτερες ενδείξεις συγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών.

Όμως στον πίνακα **Collinearity Diagnostics** και ειδικότερα οι τιμές στις στήλες **eingenvalue** και **condition index** μας επιτρέπουν να πούμε με σιγουριά ότι υπάρχει σοβαρό πρόβλημα συγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών σε ένα μοντέλο τεσσάρων διαστάσεων.

Επομένως επιβεβαιώνοντας και τα συμπεράσματα από τον προηγούμενο πίνακα καλύτερα είναι να αφαιρέσουμε την πρώτη ανεξάρτητη μεταβλητή και να τρέξουμε ξανά την παλινδρόμηση.

Model summary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,807[a]	,652	,612	10,13071	1,575

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,807[a]	,652	,626	9,94236	1,581

Όπως φαίνεται από τα αποτελέσματα των παραπάνω πινάκων πριν και μετά την αφαίρεση της πρώτης ανεξάρτητης μεταβλητής αν και λογικά έπρεπε να ήταν κάπως χειρότερα εφόσον αφαιρούμε μια μεταβλητή από το μοντέλο αυτά παρουσιάζονται και ελαφρώς καλύτερα.

Πιο συγκεκριμένα ο συντελεστής πολλαπλής συσχέτισης δεν μεταβλήθηκε το διορθωμένο R τετράγωνο αυξήθηκε επομένως μπορούμε να πούμε ότι ο βαθμός του δεύτερου και του τρίτου τεστ αξιολόγησης μπορούν να ερμηνεύσουν σε ποσοστό 65,2% την συνολική μεταβλητότητα της εξαρτημένης μεταβλητής δηλαδή του βαθμού απόδοσης των εργαζομένων. Τέλος το σταθερό λάθος μειώθηκε και ο δείκτης Durbin – Watson έγινε ελαφρώς καλύτερος. Τα παραπάνω αποτελέσματα δικαιολογούνται από την ισχυρή συγγραμμικότητα που υπήρχε μεταξύ του πρώτου και των δύο άλλων τεστ αξιολόγησης. Δοκιμάστε απλή παλινδρόμηση με εξαρτημένη μεταβλητή την χ1 και ανεξάρτητες την χ2 και την χ3

Πίνακας Anova

ANOVA[b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4994,476	3	**1664,825**	**16,221**	**,000[a]**
	Residual	2668,411	26	**102,631**		
	Total	7662,887	29			

ANOVAb						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4993,921	2	**2496,961**	**25,260**	**,000[a]**
	Residual	2668,965	27	**98,851**		
	Total	7662,887	29			

Όπως φαίνεται από τα αποτελέσματα των παραπάνω πινάκων πριν και μετά την αφαίρεση της πρώτης ανεξάρτητης μεταβλητής απορρίπτουμε και στις δύο περιπτώσεις την μηδενική **Ho ($\beta_1=\beta_2=\beta_3=0$)** ότι δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών του μοντέλου. Στην συγκεκριμένη περίπτωση μπορούμε να ισχυριστούμε ότι το δεύτερο και το τρίτο τεστ αξιολόγησης συνδυασμένα γραμμικά με την εξίσωση της γραμμικής παλινδρόμησης που θα προκύψει από τον επόμενο πίνακα συμβάλουν σημαντικά στην ερμηνεία της μεταβλητικότητας του βαθμού απόδοσης των υπαλλήλων.

Πίνακας συντελεστών

Coefficients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-28,877	19,735		-1,463	,155		
	x1	,328	4,460	,012	,073	,942	,546	1,832
	x2	3,912	1,248	,512	3,133	,004	,502	1,992
	x3	19,671	8,629	,367	2,280	,031	,516	1,938

Coefficients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-27,592	8,982		-3,072	,005		
	x2	3,946	1,136	,516	3,475	,002	,584	1,712
	x3	19,887	7,959	,371	2,499	,019	,584	1,712

Στον δεύτερο πίνακα, των συντελεστών της παλινδρόμησης, παρατηρούμε ότι και η σταθερά αλλά και οι δύο ανεξάρτητες μεταβλητές (δεύτερο και τρίτο τεστ αξιολόγησης) επιδρούν με στατιστικά σημαντικό τρόπο στην διαμόρφωση και κατά επέκταση στην ερμηνεία (πρόβλεψη) της μεταβλητικότητας της εξαρτημένης μεταβλητής δηλαδή του βαθμού απόδοσης των υπαλλήλων. Επιπλέον οι δείκτες tolerance και VIF δεν δείχνουν ιδιαίτερα προβλήματα συγγραμμικότητας.

Έλεγχος συγγραμμικότητας

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	x1	x2	x3
1	1	3,915	**1,000**	,00	,00	,00	,00
	2	,065	**7,788**	,03	,01	,57	,00
	3	,017	**15,397**	,07	,02	,28	,95
	4	,004	**33,134**	,89	,98	,15	,05

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	x2	x3
1	1	2,927	**1,000**	,00	,01	,00
	2	,058	**7,110**	,28	,65	,01
	3	,016	**13,695**	,72	,34	,99

Όπως φαίνεται από τα παραπάνω αποτελέσματα των τεστ συγγραμμικότητας μετά την αφαίρεση της πρώτης ανεξάρτητης μεταβλητής οι τιμές των στηλών Eingenvalue και Condition Index μας δείχνουν ότι και πάλι υπάρχει πρόβλημα συγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών αλλά δεν μας δημιουργεί αξεπέραστα εμπόδια επειδή οι τιμές των δεικτών είναι ελαφρώς καλύτερες από τα απαγορευτικά όρια.

Ερμηνεία αποτελεσμάτων

Coefficients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-27,592	8,982		-3,072	**,005**		
	x2	3,946	1,136	,516	3,475	**,002**	,584	1,712
	x3	19,887	7,959	,371	2,499	**,019**	,584	1,712

Όπως φαίνεται από τα παραπάνω αποτελέσματα η σταθερά και οι δύο ανεξάρτητες μεταβλητές δηλαδή το δεύτερο και το τρίτο τεστ αξιολόγησης συμβάλουν με στατιστικά σημαντικό τρόπο στην διαμόρφωση του βαθμού απόδοσης των υπαλλήλων.

Πιο συγκεκριμένα για κάθε μονάδα αύξησης στο δεύτερο τεστ αξιολόγησης ενός υπαλλήλου υποθέτοντας ότι οι μονάδες αξιολόγησης στο τρίτο τεστ παραμένουν σταθερές αναμένεται αύξηση κατά 3,94 μονάδες στην απόδοση του υπαλλήλου, ενώ μια αντίστοιχη αύξηση στο τρίτο τεστ αξιολόγησης υποθέτοντας ότι οι μονάδες αξιολόγησης στο δεύτερο τεστ παραμένουν σταθερές αναμένεται να επιφέρει 19,88 βαθμούς αύξηση στην απόδοση του υπαλλήλου. Όσο αφορά την σταθερά μπορούμε να πούμε ότι αν κάποιος υπάλληλος πάρει μηδέν στα δύο τεστ αξιολόγησης τότε α βαθμός απόδοσης του αναμένεται να είναι – 27,59. Αποτέλεσμα το οποίο δεν έχει λογική ερμηνεία οπότε χρειάζεται να κάνουμε στάθμιση

Στάθμιση

Coefficients[a]

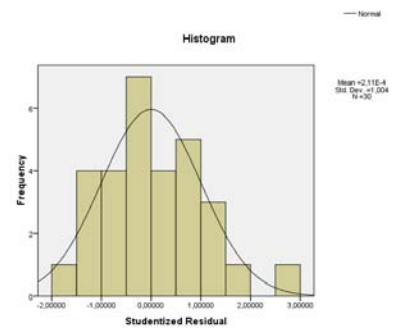
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	23,533	1,841		12,784	**,000**		
	x2a	3,946	1,136	,516	3,475	**,002**	,584	1,712
	x3a	19,887	7,959	,371	2,499	**,019**	,584	1,712

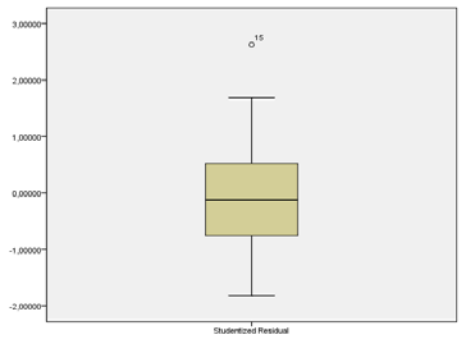
a. Dependent Variable: y

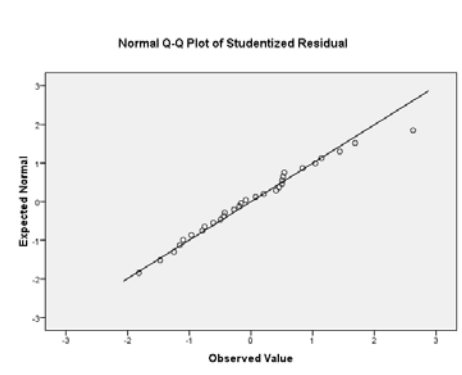
Μετά την στάθμιση των δύο ανεξάρτητων μεταβλητών που έγινε δημιουργώντας δύο νέες μεταβλητές x2a, x3a από την αφαίρεση της μέσης τιμής των αντίστοιχων μεταβλητών από τις παρατηρήσεις μπορούμε να πούμε ότι κάποιος υπάλληλος που παίρνει 5,2 στο δεύτερο τεστ και 1,4 στο τρίτο τεστ δηλαδή κυμαίνεται στον μέσο όρο του δείγματος αναμένεται να έχει 23,53 μονάδες στην αξιολόγηση της απόδοσης του.

Ορθότητα Μοντέλου (Κανονικότητα σφαλμάτων)

Από το μενού Analyze→ Descriptive Statistics→ Explore ελέγχουμε την κανονικότητα της κατανομής των καταλοίπων

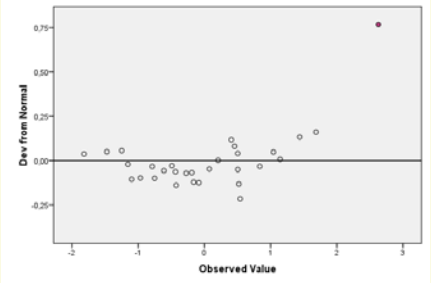






Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Studentized Residual	,096	30	,200*	,981	30	,848



Το ιστόγραμμα και τα γραφικά των studentized residuals μας δίνουν ενδείξεις ότι τα κατάλοιπα κατανέμονται κανονικά. Το τεστ για την κανονικότητα των Kolmogorov – Smirnov και των Shapiro – Wilk έχουν τιμή **μεγαλύτερη** του 0,05 και επομένως δεν μπορούμε να απορρίψουμε την υπόθεση ότι τα κατάλοιπα ακολουθούν την κανονική κατανομή. Προσοχή πρέπει να δώσουμε στην 15 παρατήρηση που φαίνεται και στο box – plot να παίρνει ακραία τιμή.

Ορθότητα Μοντέλου (Ανεξαρτησία σφαλμάτων)

Από το μενού **Analyze → non - parametric tests** εκτελούμε ένα τεστ ροών για τον έλεγχο της τυχαιότητας των σφαλμάτων. Σαν μεταβλητή επιλέγουμε την **Studentized residuals**.

Runs Test

	Studentized Residual
Test Value[a]	,0002109
Cases < Test Value	16
Cases >= Test Value	14
Total Cases	30
Number of Runs	14
Z	-,535
Asymp. Sig. (2-tailed)	**,593**

a. Mean

Με βάση το παραπάνω τεστ το οποίο εμφανίζεται στο διπλανό πίνακα κα έχει **p-value = 0,593** δεν μπορούμε να απορρίψουμε ότι τα κατάλοιπα είναι τυχαία.

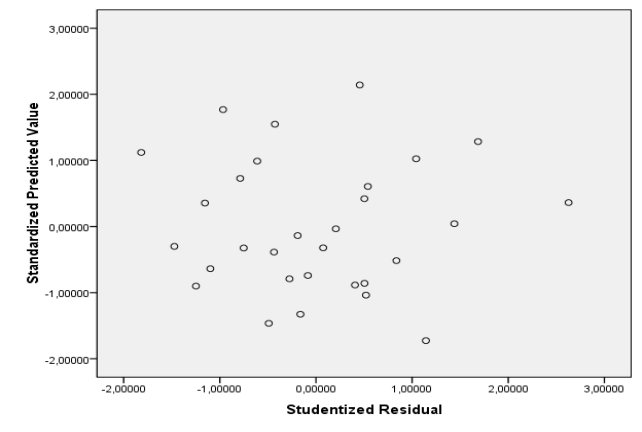
Από το πίνακα **Model Summary** παρατηρούμε επίσης ότι ο δείκτης **Durbin Watson** είναι κοντά στο **δύο** επομένως δεν έχουμε προβλήματα αυτοσυσχέτισης.

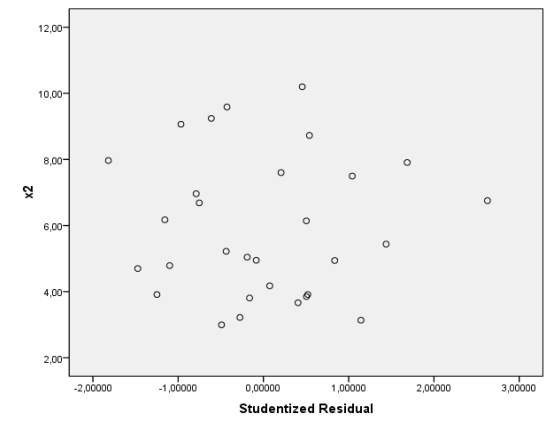
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	**,807[a]**	**,652**	**,626**	**9,94236**	**1,581**

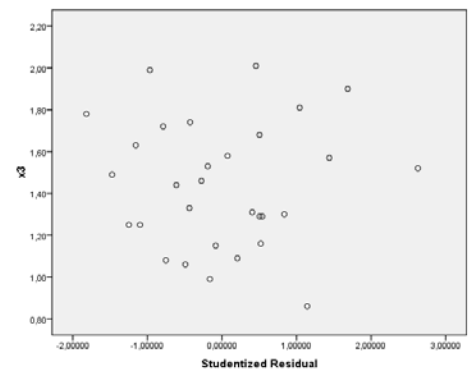
Ορθότητα Μοντέλου (Ομοσκεδαστικότητα σφαλμάτων)

Για την εκτέλεση του ελέγχου κατασκευάζουμε τα γραφήματα (scatterplot) των σημείων predicted, studentized residuals και Xi, studentized residuals





Οι παρατηρήσεις φαίνεται ότι βρίσκονται τυχαία στο επίπεδο πράγμα που υποδηλώνει ότι δεν πρέπει να υπάρχει κάποια σχέση μεταξύ των μεταβλητών (εξάλλου με τόσες λίγες παρατηρήσεις δεν είναι εύκολο να ανακαλύψουμε κάτι τέτοιο).



Ορθότητα Μοντέλου (Ακραίες τιμές)

Παρατηρώντας την στήλη των Studentized residuals επιβεβαιώνουμε ότι στην παρατήρηση 15 παρατηρούνται έκτροπες τιμές (>2) επομένως πρέπει να ελέγξουμε τις τιμές σε αυτή την παρατήρηση

Για να εξετάσουμε αν υπάρχουν παρατηρήσεις που έχουν μεγάλη «*επιρροή*» στο μοντέλο ελέγχουμε ποιες έχουν centered leverage > 5/n = 5/30 = 0,166. Βλέπουμε ότι ή έκτη παρατήρηση (0,18) έχει μεγάλη επιρροή στο μοντέλο καμία παρατήρηση καθώς και 7 έβδομη παρατήρηση πλησιάζει το όριο (τέτοιες παρατηρήσεις πρέπει να λαμβάνονται με προσοχή).

	Studentized residuals	Centered leverage values		Studentized residuals	Centered leverage values
1	1,143583	0,127219	16	-0,427175	0,102246
2	-0,190347	0,024829	17	0,520057	0,037219
3	1,438917	0,021415	18	-1,473356	0,028625
4	1,042051	0,050854	19	-1,155515	0,018204
5	-0,162043	0,076598	20	-0,96646	0,117871
6	0,206849	0,184196	21	-0,084177	0,033314
7	0,539741	0,165859	22	-0,436757	0,005303
8	-1,817987	0,045992	23	-0,611395	0,142893
9	0,504618	0,034403	24	0,455301	0,158073
10	0,074663	0,081747	25	0,406191	0,043406
11	0,503665	0,031202	26	-0,27556	0,102717
12	1,686304	0,078671	27	-0,788445	0,02966
13	-0,490942	0,074037	28	-1,248561	0,031423
14	0,836304	0,009222	29	-0,751934	0,122911
15	2,627527	0,005012	30	-1,098789	0,014877

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.

Λίγα λόγια για την επιλογή μοντέλου

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.

- 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).

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 ψευδομεταβλητών (dummy variables). Για κάθε κατηγορία της ποιοτική μεταβλητή φτιάχνεται μία ψευδομεταβλητή. Κάθε ψευδομεταβλητή 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 ψευδομεταβλητές όσες ο αριθμός των κατηγοριών της μεταβλητής μείον 1. Η ψευδομεταβλητή που δεν εισάγεται στο μοντέλο αποτελεί το επίπεδο αναφοράς (reference level).

ΤΕΛΟΣ