Ανάλυση Συνδιακύμανσης – Analysis of Covariance

Η ανάλυση συνδιακύμανσης είναι μία άλλη τεχνική για να βελτιώσουμε την ακρίβεια της προσέγγισης του μοντέλου μας στο πείραμα. Ας υποθέσουμε ότι σ' ένα πείραμα εκτός από την εξαρτημένη μεταβλητή y, υπάρχει και μία άλλη μεταβλητή x, η οποία συνδέεται γραμμικά με την y. Η μεταβλητή x δε μπορεί να ελεγχθεί (δεν μπορεί να σταθεροποιηθεί) αλλά μπορεί να μετρηθεί μαζί με την y. Με την ανάλυση διακύμανσης καταφέρνουμε να ελέγξουμε την επίδραση της x πάνω στην εξαρτημένη μεταβλητή y.

Π.χ.: Έστω ότι θέλουμε να ελέγξουμε την ανθεκτικότητα του νήματος που παράγεται από τρεις διαφορετικές μηχανές. Όμως στην ανθεκτικότητα του νήματος (που είναι η απαντητική μεταβλητή y) επιδρά και η διάμετρός του (που είναι η μεταβλητή x), την οποία εμείς δεν μπορούμε να σταθεροποιήσουμε, δηλαδή κάθε μηχανή να βγάζει συνεχώς νήμα της ίδιας διαμέτρου.

Το παρακάτω μοντέλο θα περιγράφει το πρόβλημά μας:

$$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) + \varepsilon_{ij} \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases}$$

a ο αριθμός των μηχανών
n ο αριθμός των δειγμάτων που παίρνω από κάθε μηχανή

όπου: β : συντελεστής παλινδρόμησης (που είναι ίδιος για όλα τα επίπεδα)
x_{ij} : η διάμετρος του νήματος της κάθε μονάδας του δείγματος
$\bar{x}_{\bullet\bullet}$: η over all mean των διαμέτρων

Για να περιγράψουμε την ανάλυση συνδιακύμανσης, εισάγουμε τους παρακάτω συμβολισμούς:

- $S_{yy} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{an}$
- $S_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - \frac{x_{\bullet\bullet}^2}{an}$
- $S_{xy} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right)\left(y_{ij} - \bar{y}_{\bullet\bullet}\right) = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}y_{ij} - \frac{(x_{\bullet\bullet})(y_{\bullet\bullet})}{an}$
- $T_{yy} = n\sum_{i=1}^{a}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2 = \frac{1}{n}\sum_{i=1}^{a} y_{i\bullet}^2 - \frac{y_{\bullet\bullet}^2}{an}$
- $T_{xx} = n\sum_{i=1}^{a}\left(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}\right)^2 = \frac{1}{n}\sum_{i=1}^{a} x_{i\bullet}^2 - \frac{x_{\bullet\bullet}^2}{an}$

- $T_{xy} = n\sum_{i=1}^{a}\left(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}\right)\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right) = \frac{1}{n}\sum_{i=1}^{a}(x_{i\bullet})(y_{i\bullet}) - \frac{(x_{\bullet\bullet})(y_{\bullet\bullet})}{an}$

- $E_{yy} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2 = S_{yy} - T_{yy}$

- $E_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2 = S_{xx} - T_{xx}$

- $E_{xy} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\bullet}\right)\left(y_{ij} - \bar{y}_{i\bullet}\right) = S_{xy} - T_{xy}$

Εκτίμηση των Παραμέτρων του Μοντέλου (Μέθοδος Ελαχίστων Τετραγώνων)

$$L = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[y_{ij} - \mu - \tau_i - \beta\left(x_{ij} - \bar{x}_{\bullet\bullet}\right)\right]^2$$

μ: $an\hat{\mu} + n\sum_{i=1}^{a}\hat{\tau}_i = y_{\bullet\bullet}$ (I)

τ_i: $n\hat{\mu} + n\hat{\tau}_i + \hat{\beta}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) = y_{i\bullet}$ (II)

β: $\sum_{i=1}^{a}\hat{\tau}_i\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) + \hat{\beta}\cdot S_{xx} = S_{xy}$ (III)

Ισχύει ότι:

$$\sum_{i=1}^{a}\hat{\tau}_i = 0 \quad \text{και} \quad \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) = 0$$

Επίσης οι εξισώσεις (I) και (II) είναι γραμμικά εξαρτημένες

Οι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων θα είναι:

$$\hat{\mu} = \bar{y}_{\bullet\bullet} \qquad \hat{\tau}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} - \hat{\beta}\left(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}\right)$$

$$\hat{\beta} = \frac{S_{xy} - T_{xy}}{S_{xx} - T_{xx}} = \frac{E_{xy}}{E_{xx}}$$

E_{xy} : Τα λάθη που οφείλονται στη σχέση του y με το x

E_{xx} :Τα λάθη που οφείλονται στη μεταβλητή x

F – Test

Το μοντέλο μας είναι το εξής:

$$y_{ij} = \mu' + \tau_i + \beta x_{ij} + \varepsilon_{ij} \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases}$$

όπου: $\mu' = \mu - \beta\bar{x}_{\bullet\bullet}$

Θα έχουμε:

$$SS_E = E_{yy} - (E_{xy})^2 / E_{xx} \quad \text{και} \quad MS_E = \frac{SS_E}{a(n-1)-1}$$

Αν υποθέσουμε ότι δεν έχουμε καμία επίδραση από τα επίπεδα $\Leftrightarrow \tau_i = 0$, θα έχουμε:

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}_{\bullet\bullet}) + \varepsilon_{ij} \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases}$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \text{και} \quad \hat{\mu} = \bar{y}_{\bullet\bullet}$$

Αντίστοιχα θα έχουμε:

$$SS'_E = S_{yy} - (S_{xy})^2 / S_{xx}$$

Αποδεικνύεται ότι:

$$MS_E \sim X^2_{\alpha,a(n-1)-1} \quad \text{και} \quad \frac{SS'_E - SS_E}{a-1} \sim X^2_{\alpha,a-1}$$

Για να ελέγξω την αρχική υπόθεση $H_0 : \tau_i = 0$, χρησιμοποιώ το F – Test, δηλαδή απορρίπτω την H_0 εάν:

$$F_0 > F_{\alpha,a-1,a(n-1)-1}$$

όπου:

$$F_0 = \frac{(SS'_E - SS_E)/(a-1)}{MS_E}$$

Παράδειγμα

Έστω ότι έχουμε το πρόβλημα που αναφέραμε στην αρχή της παραγράφου, όπου έχουμε να ελέγξουμε την ανθεκτικότητα του νήματος που παράγεται από τρεις διαφορετικές μηχανές και στην οποία επιδρά και η διάμετρος του νήματος. Από μετρήσεις προκύπτουν τα παρακάτω αριθμητικά δεδομένα:

ΜΗΧΑΝΗ 1		**ΜΗΧΑΝΗ 2**		**ΜΗΧΑΝΗ 3**	
y	**x**	**y**	**x**	**y**	**x**
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15

Υποθέτουμε ότι η διάμετρος του νήματος συνδέεται γραμμικά με τη διάμετρό του. Το πρόβλημά μας θα περιγράφεται από το παρακάτω μοντέλο:

$$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) + \varepsilon_{ij} \begin{cases} i = 1,2,3 \\ j = 1,2,3,4,5 \end{cases}$$

όπου i ο αριθμός των μηχανών και j ο αριθμός των δειγμάτων που παίρνουμε από κάθε μηχανή
Για να προχωρήσουμε στην ανάλυση συνδιακύμανσης υπολογίζουμε τις σχέσεις:

$S_{yy} = 346,40$ $\quad T_{yy} = 140,40$ $\quad E_{yy} = 206,00$

$S_{xx} = 261,73$ $\quad T_{xx} = 66,13$ $\quad E_{xx} = 195,60$

$S_{xy} = 282,60$ $\quad T_{xy} = 96,00$ $\quad E_{xy} = 186,60$

Επίσης θα έχουμε:

$$SS_E = 27.99 \quad \text{και} \quad MS_E = 2,54$$

όπως και

$$SS'_E = 41,27$$

Για να ελέγξουμε την υπόθεση αν οι μηχανές παράγουν νήμα διαφορετικής ανθεκτικότητας ($H_0 : \tau_i = 0$) χρησιμοποιούμε F – Test:

$$F_0 = \frac{13{,}28/2}{27{,}99/11} = 2{,}61 \quad \text{και} \quad p - \text{value} = 0{,}1181$$

Άρα δεν έχουμε ισχυρούς λόγους για να πιστεύουμε ότι οι μηχανές παράγουν νήμα διαφορετικής ανθεκτικότητας.

Προϋποθέσεις του μοντέλου

Ο έλεγχος των υποθέσεων του μοντέλου, οι απαραίτητες δηλαδή συνθήκες που πρέπει να ικανοποιούνται έτσι ώστε να μπορεί να χρησιμοποιηθεί αυτό για στατιστική συμπερασματολογία, γίνεται μέσω των καταλοίπων. Τα κατάλοιπα όπως γνωρίζουμε ορίζονται ως

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

ενώ οι εκτιμηθείσες τιμές

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}(x_{ij} - \bar{x}_{..}) = \bar{y}_{..} + [\bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})] + \hat{\beta}(x_{ij} - \bar{x}_{..}) =$$

$$= \bar{y}_{i.} + \hat{\beta}(x_{ij} - \bar{x}_{i.})$$

Έτσι, τα κατάλοιπα γίνονται

$$e_{ij} = y_{ij} - \bar{y}_{i.} - \hat{\beta}(x_{ij} - \bar{x}_{i.}) .$$

Στο προηγούμενο παράδειγμα με τη διάμετρο των νημάτων θα έχουμε για το κατάλοιπο της πρώτης παρατήρησης της πρώτης μηχανής

$$e_{11} = y_{11} - \bar{y}_{1.} - \hat{\beta}(x_{11} - \bar{x}_{1.}) = 36 - 41.4 - (0.9540)(20 - 25.2) =$$

$$= 36 - 36.4392 = -0.4392$$

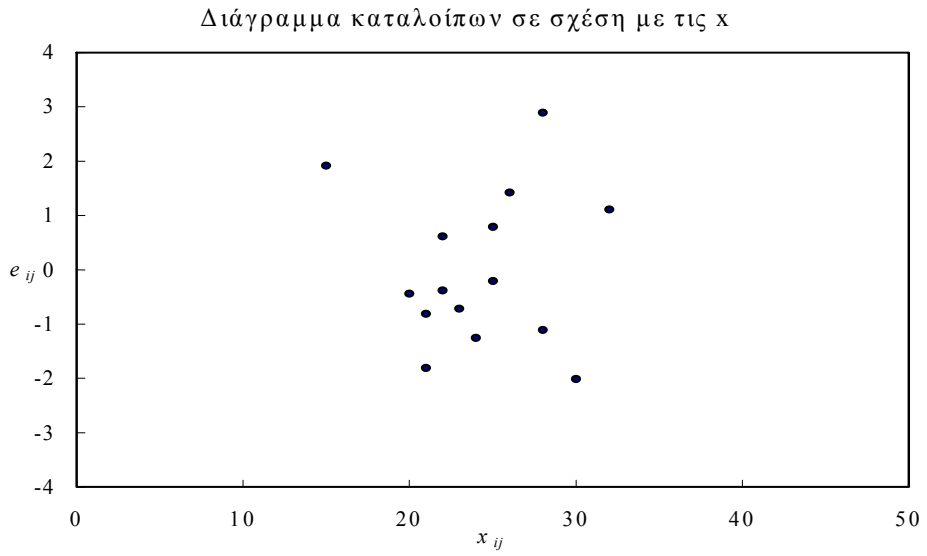
Με παρόμοιο τρόπο υπολογίζονται και τα υπόλοιπα κατάλοιπα και δίνονται παρακάτω μαζί με τις παρατηρηθείσες και τις εκτιμηθείσες τιμές.

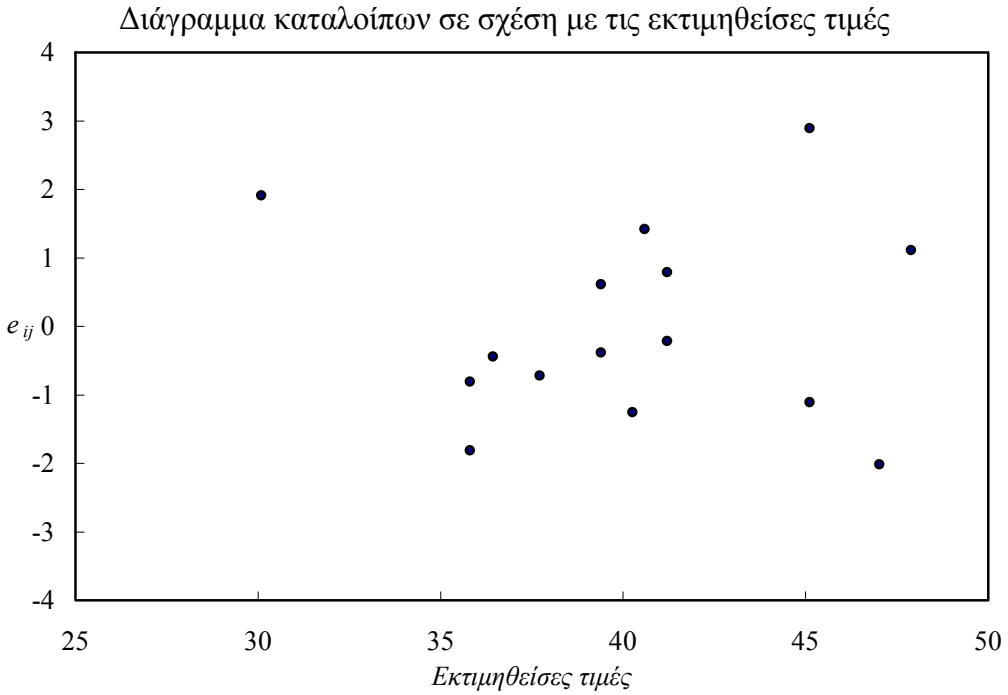
Παρατηρηθείσες τιμές y_{ij}	Εκτιμηθείσες τιμές $\hat{y}_{ij}$	Κατάλοιπα e_{ij}
36	36,4393	-0,4393
41	41,2092	-0,2092
39	40,2552	-1,2552
42	41,2092	0,7908
49	47,8871	1,1129
40	39,384	0,616
48	45,108	2,892
39	39,384	-0,384
45	47,016	-2,016
44	45,108	-1,108
35	35,8092	-0,8092
37	37,7172	-0,7172
42	40,5791	1,4209
34	35,8092	-1,8092
32	30,0853	1,9147

Με βάση τον παραπάνω πίνακα κατασκευάζονται και τα παρακάτω διαγράμματα. Το πρώτο από αυτά, είναι το διάγραμμα των καταλοίπων σε σχέση με τις εκτιμηθείσες τιμές (residuals vs fitted values)

Παρατηρώντας το διάγραμμα βλέπουμε πως τα κατάλοιπα δεν έχουν μια προκαθορισμένη σειρά σε σχέση με τις εκτιμηθείσες τιμές (είναι τυχαία

τοποθετημένα και σχηματίζουν ένα «σύννεφο») το οποίο μας οδηγεί στο συμπέρασμα ότι τα κατάλοιπα αυτά είναι ανεξάρτητα μεταξύ τους. Στο αντίθετο συμπέρασμα θα καταλήγαμε βέβαια αν τα σημεία στο χώρο του διαγράμματος ακολουθούσαν μια συγκεκριμένη πορεία ή είχαν μια προσδιορισμένη μορφή, όπως για παράδειγμα αν σχημάτιζαν μια νοητή ευθεία ή έστω μια άλλη μη τυχαία διάταξη. Περίπου στο ίδιο συμπέρασμα καταλήγουμε και από το διάγραμμα των καταλοίπων σε σχέση με τις παρατηρηθείσες





τιμές το οποίο παριστάνεται μετά το διάγραμμα καταλοίπων σε σχέση με τις $\hat{y}_{ij}$.

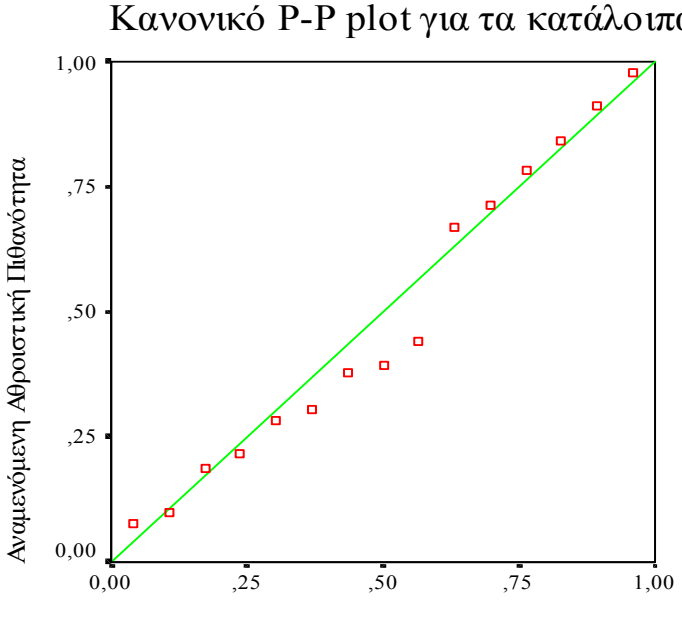
Αφού λοιπόν και το δεύτερο γράφημα δείχνει ένα «σμήνος» από κουκίδες τυχαία τοποθετημένες στο επίπεδο αποκλείεται και εδώ η περίπτωση εξάρτησης μεταξύ των καταλοίπων.

Στο επόμενο διάγραμμα έχουμε τα κατάλοιπα σε σχέση με τις τρεις μηχανές που χρησιμοποιήθηκαν κατά σειρά.

Με τα διαγράμματα που παρουσιάζουν τα κατάλοιπα σε σχέση με τις μηχανές ελέγχουμε την ύπαρξη ετεροσκεδαστικότητας ή όχι. Συγκεκριμένα, αν το εύρος των καταλοίπων αυξάνεται ή μειώνεται σε κάθε μηχανή σταδιακά τότε αυτό αποτελεί ένδειξη ύπαρξης ετεροσκεδαστικότητας. Στο συγκεκριμένο παράδειγμα, υποθέτουμε πως η διακύμανση στις τρεις μηχανές είναι περίπου η ίδια ότι έχουμε δηλαδή ομοσκεδαστικότητα μιας που το εύρος το καταλοίπων (το «άνοιγμα» από το υψηλότερο έως το χαμηλότερο σημείο) από την πρώτη έως και την τρίτη μηχανή είναι περίπου το ίδιο. Σε αυτό το σημείο βέβαια θα πρέπει να προσθέσουμε πως η σειρά των μηχανών φυσικά μπορεί να αλλάξει μιας που δεν έχει προκαθορισθεί και αυτό άλλωστε δεν επηρέασε τη μέχρι τώρα πορεία. Ακόμη όμως και αν αλλάξει αυτό δεν μπορεί και δεν πρέπει να αλλάζει τα συμπεράσματα μας αφού ένα σχετικά πολύ μεγαλύτερο εύρος τιμών θα πρέπει να προβληματίσει τον ερευνητή είτε αυτή η μηχανή είναι στη μέση είτε στην αρχή είτε στο τέλος.

Το σημαντικότερο ίσως απ' όλα τα διαγράμματα που κατασκευάζουμε για τα κατάλοιπα είναι το P-P plot ή το διάγραμμα ελέγχου κανονικότητας. Όπως λεει και το όνομα του είναι το διάγραμμα που μας δείχνει κατά πόσο τα κατάλοιπα της εκτιμηθείσας εξίσωσης ακολουθούν την κανονική κατανομή. Παρακάτω έχουμε ένα P-P plot όπως αυτό κατασκευάζεται από το στατιστικό πακέτο SPSS.

Σύμφωνα με το διάγραμμα αυτό, για να ακολουθούν τα κατάλοιπα (κουκκίδες) κανονική κατανομή θα πρέπει να είναι όσο το δυνατόν πιο κοντά στη κεντρική γραμμή, αν είναι δυνατόν να συμπίπτουν με αυτή για να έχουμε και τέλεια



κανονικότητα. Στον οριζόντιο άξονα του διαγράμματος αναγράφονται οι παρατηρηθείσες αθροιστικές πιθανότητες ενώ στον κάθετο άξονα είναι οι αναμενόμενες αθροιστικές πιθανότητες. Στη δική μας περίπτωση, μπορούμε με ασφάλεια να υποθέσουμε πως τα κατάλοιπα μας ακολουθούν την κανονική κατανομή αφού αυτά κινούνται πολύ κοντά στη κεντρική γραμμή.

Έτσι, με όλα αυτά τα διαγράμματα μπορέσαμε να διαπιστώσουμε αν ικανοποιούνται οι προϋποθέσεις του μοντέλου και οι οποίες μας επιτρέπουν να το χρησιμοποιήσουμε για στατιστική συμπερασματολογία. Θα πρέπει ωστόσο να αναφέρουμε πως καλό είναι να μη βασιζόμαστε αποκλειστικά και μόνο στα διαγράμματα για να αποφασίσουμε αν ένα μοντέλο ικανοποιεί μια συνθήκη μιας που χωρά μεγάλος βαθμός υποκειμενικότητας στις εκτιμήσεις καθένα μας. Έτσι το σωστότερο θα ήταν να πραγματοποιούμε ορισμένα τεστ και σύμφωνα με αυτά να κρίνουμε. Εν τούτοις τα διαγράμματα χρησιμοποιούνται ευρύτατα μιας που έχουν ένα πιο γρήγορο και εν μέρει αξιόπιστο τρόπο αξιολόγησης παρόλο που έγκεινται στον εκάστοτε ερευνητή αν θα δεχθεί ή όχι την ισχύ μιας προϋπόθεσης των καταλοίπων.

Είναι επίσης πολύ ενδιαφέρον να σημειώσουμε πως αν το πρόβλημα της αυτοσυσχέτισης δεν είχε εντοπισθεί και αγνοούσαμε την επίδραση της τ.μ. x στην y τότε ως λογικό αντί για ανάλυση συνδιακύμανσης θα κάναμε ανάλυση διακύμανσης με ένα παράγοντα. Τότε, όπως αποδεικνύεται, θα καταλήγαμε σε λάθος συμπέρασμα. Συγκεκριμένα, με την ανάλυση διακύμανσης με ένα παράγοντα καταλήγουμε στο συμπέρασμα ότι οι μηχανές θα διαφέρανε στατιστικά σημαντικά μεταξύ τους σύμφωνα και με τον παρακάτω πίνακα.

Source	SS	D.F.	M.S.	F_0	p-value
Μηχανές	140,40	2	70,20	4,09	0,0442
Κατάλοιπα	206,00	12	17,17		
Σύνολο	346,40	14			

Όπως αποδείξαμε όμως εμείς προηγουμένως με την ανάλυση συνδιακύμανσης οι μηχανές δε διαφέρουν στατιστικά σημαντικά. Καταλήξαμε επομένως σε δύο αντίθετα αποτελέσματα με τις δύο αναλύσεις. Επομένως είναι σημαντικό να αναγνωρίζουμε την επίδραση των nuisance factors ή αλλιώς των «ενοχλητικών» παραγόντων όπως είναι για το παράδειγμα μας η διάμετρος του νήματος η οποία αλλάζει την αντοχή του σε κάθε σημείο του.