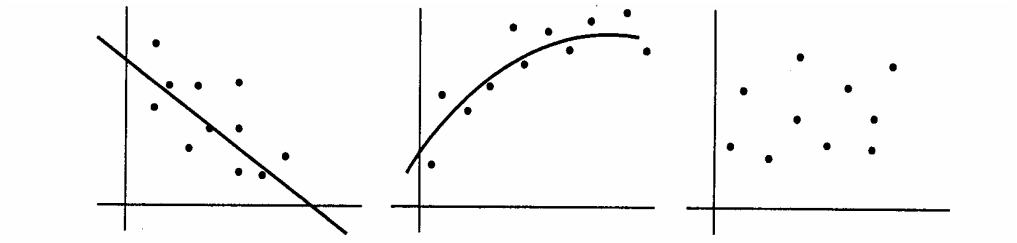
ΠΑΛΙΝΔΡΟΜΗΣΗ



ΓΡΑΜΜΙΚΟ ΜΗ ΓΡΑΜΜΙΚΟ ΔΕΝ ΥΠΑΡΧΕΙ ΣΧΕΣΗ

Απλή Παλινδρόμηση

$$Y = a + bx + e$$

(Όγκος πωλήσεων = α +b έξοδα διαφήμισης +e)

Εκτίμηση Απλής Παλινδρόμησης

$$\hat{Y} = \hat{a} + \hat{b}x$$

(Όγκος πωλήσεων = α +b έξοδα διαφήμησης +e)

α= εκτίμηση της τεταγμένης για χ=0

b= εκτίμηση της κλίσης

Y=εξαρτημένη μεταβλητή

X= ανεξάρτητη μεταβλητή ή ερμηνεύσιμη

Η μεταβλητή X ερμηνεύει τις μεταβολές της Y που εκφράζεται σαν ποσοστό επί της %

.

Αν έχουμε πλήρη συσχέτισης τότε έχουμε 100%. Δηλ., Οι τιμές Y ταυτίζονται με τις X πάνω σε μια ευθεία.

Το σφάλμα πρόβλεψης ονομάζεται κατάλοιπο.

Δίνεται σαν η διαφορά των Y και $\hat{Y}$**. Δηλαδή,**

$$Y - \hat{Y}$$

Σφάλμα εκτίμησης

Άθροισμα τετραγώνων των καταλοίπων διαιρούμενων με n-2.

ΠΟΛΥΔΙΑΣΤΑΤΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

$$Y = a + b_1 x_1 + \ldots + b_k x_k + e$$

ΠΑΡΑΔΟΧΕΣ ΜΟΝΤΕΛΟΥ

1) **Σφάλματα κανονικά.**
2) **Μέσο σφάλμα 0**
3) **Διακύμανση σφάλματος σ^2**
4) **Τα σφάλματα ανεξάρτητα.**

Το άθροισμα των τετραγώνων των συνολικών παρατηρήσεων από το μέσο, $\Sigma (Y_i - \overline{Y})^2$, καλείται **συνολικό άθροισμα τετραγώνων** ή **TSS** (Total Sum of Squares). Μαθηματικά αυτό ισούται με : $\Sigma (\hat{Y} - \overline{Y})^2 + \Sigma (Y - \hat{Y})^2$. Δηλαδή με το **άθροισμα τετραγώνων της παλινδρομήσεως** (**SSR**) και με το **άθροισμα τετραγώνων των καταλοίπων** (SSE).

Συνεπώς ισχύει η σχέση: TSS = SSR + SSE

$R^{2=}$ = **Ερμηνεύσιμη μεταβολή/Ολική μεταβολή =1-Άθροισμα τετραγώνων καταλοίπων /Ολική μεταβολή Ή**

$$\mathbf{R^2}=\frac{Άθροισμα\ τετραγώνων\ παλινδρόμησης}{Συνολικό\ άθροισμα\ τετραγώνων}=\frac{SSR}{TSS}$$

Διορθωμένος πολυδιάστατος συντελεστής παλινδρόμησης
(Adjusted R-Squared)

$$\mathbf{R\text{-}sq(adj)= R^2_A= 1-(1-R^2)(}\frac{n-1}{n-k-1}\mathbf{)}$$

Όπου: n= Μέγεθος δείγματος

k= Αριθμός ανεξάρτητων μεταβλητών μέσα στο μοντέλο

Τυπικό σφάλμα εκτίμησης

Εκτίμηση για την τυπική απόκλιση του μοντέλου

$$\mathbf{s_\varepsilon}=\sqrt{\frac{SSE}{n-k-1}}=\sqrt{MSE}$$

όπου SSE = Άθροισμα των τετραγώνων των καταλοίπων
n = Μέγεθος δείγματος
k = Αριθμός ανεξάρτητων μεταβλητών

ΕΚΤΙΜΗΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

$$\hat{Y}=\hat{a}+\hat{b_1}x_1+...+\hat{b_k}x_k$$

I) Προσδιορισμός Μοντέλου : Εξαρτημένη -Ανεξάρτητη Μεταβλητή-Γραμμική Παλινδρόμηση.

II). ***Κατασκευή Μοντέλου: Υπολογισμοί Εκτιμήσεων του Μοντέλου.***

1). **Είναι το μοντέλο στατιστικά σημαντικό ;** .

Η μηδενική και η εναλλακτική υπόθεση, ελέγχονται σε επίπεδο σημαντικότητας α(=0.05 ή 0,01 ή…). Ελέγχω την H : $\beta_1 = \beta_2 = \beta_{3.....} = \beta k = 0$ έναντι της

Α: Τουλάχιστον ένα β_i δεν είναι ίσο με το 0 σε επίπεδο σημαντικότητας α.

Αν ισχύει η μηδενική υπόθεση και όλοι οι συντελεστές είναι 0, τότε το μοντέλο της παλινδρόμησης δεν είναι ικανό να προβλέψει ή να περιγράψει.

Ο έλεγχος της F, είναι μια μέθοδος με την οποία ελέγχουμε αν το μοντέλο παλινδρόμησης μπορεί να εξηγήσει ένα σημαντικό μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής. Ο έλεγχος της στατιστικής F για την πολυδιάστατη παλινδρόμηση είναι

Έλεγχος της στατιστικής F

$$\mathbf{F} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE}$$

Όπου: Πολυδιάστατος συντελεστής παλινδρόμησης (R^2)

$$\mathbf{R^2} = \frac{Άθροισμα\ τετραγώνων\ παλινδρόμησης}{Συνολικό\ άθροισμα\ τετραγώνων} = \frac{SSR}{TSS}$$

n= Αριθμός δεδομένων

k= Αριθμός ανεξάρτητων μεταβλητών

Βαθμοί ελευθερίας = $D_1 = k$ και $D_2 = n-k-1$

Από την έξοδο του προγράμματος

Analysis of variance αν δίνει p- τιμή F μικρό τότε δέχομαι την Α.

2). **Είναι οι μεταβλητές από μόνες τους σημαντικές;**

H: $\beta_i = 0$, δεδομένου ότι όλες οι άλλες μεταβλητές είναι ήδη στο μοντέλο

A : $\beta_i \neq 0$, για κάθε i

Ο έλεγχος των υποθέσεων, μπορεί να γίνει χρησιμοποιώντας τον έλεγχο της t.

Έλεγχος t

$$t = \frac{b_i - 0}{s_{b_i}}$$

όπου: b_i: Συντελεστής κλίσης δείγματος για την ανεξάρτητη i μεταβλητή
s_{bi}: Εκτίμηση τυπικού σφάλματος για τον i συντελεστή κλίσης του δείγματος
Βαθμοί ελευθερίας = n-k-1

Από την έξοδο του προγράμματος, αν έχουμε p- τιμή t μικρές δέχομαι την Η.

3). **Υπάρχει Πολυσυγραμμικότητα;** • Συσχετίσεις μεταξύ ανεξαρτήτων μεταβλητών μικρές, αν είναι δυνατόν μηδενικές .

- • Όταν υπάρχει πολυσυγγραμμικότητα, οι συντελεστές β των ανεξάρτητων μεταβλητών είναι ευμετάβλητοι, και ακόμα και το πρόσημό τους είναι πιθανόν να αλλάζει όταν συμπεριλαμβάνονται διαφορετικές μεταβλητές. Επίσης οι τιμές P μπορεί να διογκωθούν, και αυτό θα έχει σαν αποτέλεσμα να μην απορριφθεί η μηδενική υπόθεση, ενώ θα έπρεπε να απορριφθεί στην πραγματοποίηση του ελέγχου για την σημαντικότητα του μοντέλου.

• • • ΕΠΙΠΡΟΣΘΕΤΑ ,
Διαφορά πληθωριστικού παράγοντα (VIF)

$$\mathbf{VIF} = \frac{1}{1 - R_j^2}$$

όπου R^2_j = Συντελεστής παλινδρόμησης όταν η j ανεξάρτητη μεταβλητή στρέφεται εναντίον των υπολοίπων k-1 ανεξάρτητων μεταβλητών.

Εάν VIF < 5 για μια συγκεκριμένη μεταβλητή, η πολυσυγγραμικότητα δεν θεωρείται πρόβλημα για αυτή την μεταβλητή. Αν VIF ≥ 5 μας δείχνει ότι η συσχέτιση ανάμεσα στις ανεξάρτητες μεταβλητές είναι πολύ μεγάλη και πρέπει να αντιμετωπιστεί αφαιρώντας μεταβλητές από το μοντέλο.

4). Ανάλυση καταλοίπων (Επιβεβαίωση παραδοχών)

1. **Σφάλματα κανονικά → Ιστόγραμμα καταλοίπων συμμετρικό .**

2. **Μέσο σφάλμα 0 → Άθροισμα καταλοίπων μηδέν**
3. **Διακύμανση σφάλματος σ^2 σταθερή → Διάγραμμα καταλοίπων έναντι τιμών της X (ή $\widehat{Y}$) γεμίζει ομοιογενώς το διάγραμμα.**
4. **Τα σφάλματα ανεξάρτητα. →**

Durbin-Watson. Το κριτήριο αυτό βασίζεται στην κατανομή δειγματοληψίας της στατιστικής :

$$d = \frac{\sum_{t=2} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1} \hat{e}_t^{\,2}}$$

που είναι συνήθως γνωστή ως στατιστική Durbin-Watson d στατιστική (Durbin-Watson d statistic).Οι τιμές που μπορεί να πάρει, η στατιστική d, κυμαίνονται ανάμεσα στην τιμή μηδέν και στην τιμή τέσσερα.

Όταν 0 < d < 2, τότε υπάρχει κάποιος βαθμός θετικής αυτοσυσχετίσεως, ενώ όταν 2 < d

< 4, τότε υπάρχει κάποιος βαθμός αρνητικής αυτοσυσχετίσεως.

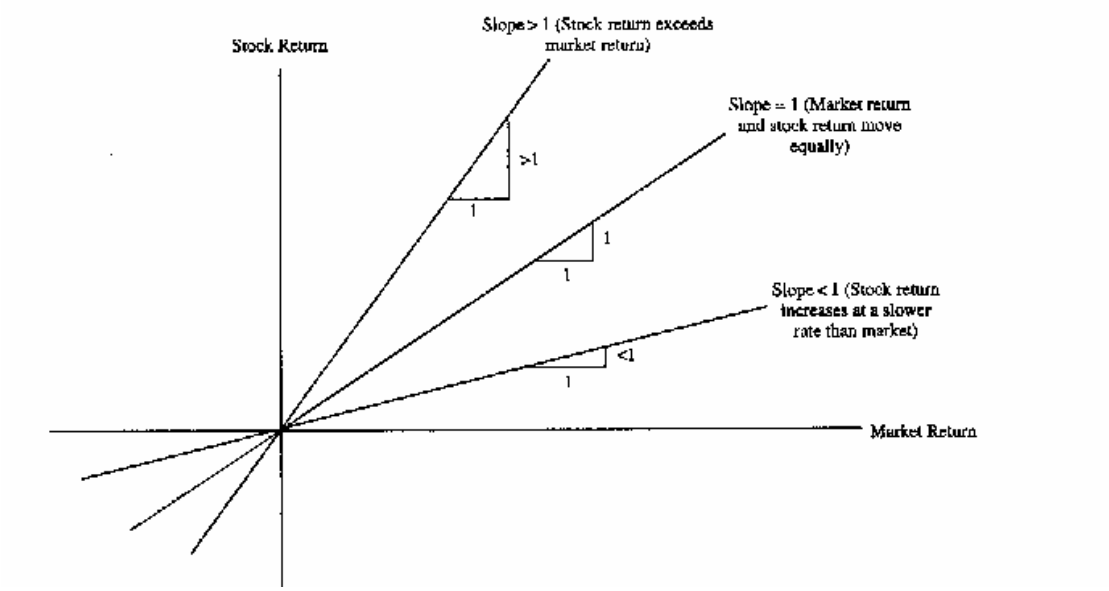
ΕΦΑΡΜΟΓΗ

Εφαρμογή της Ανάλυσης της Παλινδρόμησης στον Επενδυτικό Κίνδυνο

Οι επενδύσεις στο χρηματιστήριο είναι ελκυστικές σε όλους. Παρόλα αυτά, οι χρηματιστηριακές επενδύσεις μεταφέρουν και το στοιχείο του κινδύνου. Ο κίνδυνος που αντιστοιχεί σε κάθε μετοχή, μπορεί να μετρηθεί με δύο τρόπους. Ο πρώτος είναι ο **συστηματικός κίνδυνος** (systematic risk), που εξηγεί την μεταβλητότητα που δημιουργείται στην αξία της μετοχής, καθώς η αγορά κινείται πάνω ή κάτω, η αξία τείνει να κινείται και αυτή προς την ίδια κατεύθυνση. Ο δείκτης Standar & Poor's (S&P) 500 είναι το πιο συνηθισμένο μέτρο που χρησιμοποιείται στην αγορά. Ο δεύτερος τύπος κινδύνου καλείται **ειδικός κίνδυνος** (specific risk), και δείχνει την μεταβλητότητα που οφείλεται σε άλλου παράγοντες , όπως είναι η δυνατότητα αποδοχών της εταιρείας , οι στρατηγικές αποκτήσεων και τα λοιπά. Ο ειδικός κίνδυνος υπολογίζεται από το τυπικό σφάλμα της εκτίμησης.

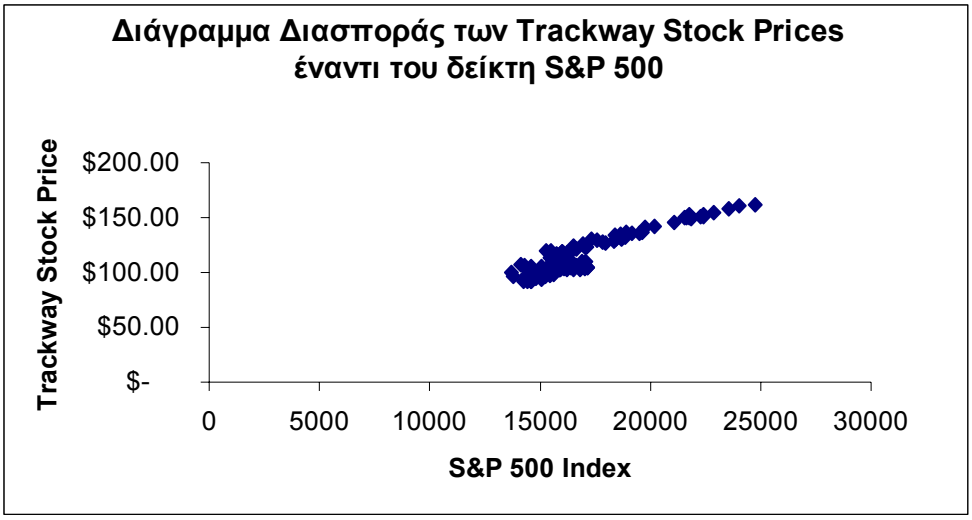
Ο συστηματικός κίνδυνος χαρακτηρίζεται από ένα μέτρο που καλείται **βήτα** (beta) . Οι τυποποιημένες τιμές, γνωστές και ως beta τιμές, προκύπτουν σαν εκτιμηθείσες τιμές των παραμέτρων του μοντέλου, αφού πρώτα μετατρέψουμε τις τιμές της εξαρτημένης και ανεξάρτητης μεταβλητής σε τυποποιημένες. Στην απλή γραμμική παλινδρόμηση, η beta τιμή της κλίσης ισούται με τον συντελεστή συσχέτισης των δύο μεταβλητών. Μια αξία beta που ισοδυναμεί με 1,0 δηλώνει ότι η συγκεκριμένη μετοχή θα ακολουθεί τις μετακινήσεις της αγοράς, ενώ ένα beta μικρότερο από 1,0 δείχνει ότι η μετοχή είναι λιγότερο ασταθής από τη αγορά . Ένα beta μεγαλύτερο από 1,0 δείχνει ότι η μετοχή έχει μεγαλύτερη διακύμανση από την αγορά. Ακόμα, μετοχές με μεγαλύτερες τιμές beta είναι περισσότερο επικίνδυνες από αυτές με χαμηλότερες τιμές beta.

Οι τιμές beta μπορούν να υπολογισθούν αναπτύσσοντας ένα μοντέλο παλινδρόμησης με τις αποδόσεις των μετοχών (εξαρτημένη μεταβλητή) έναντι του μέσου όρου των αποδόσεων της αγοράς (ανεξάρτητη μεταβλητή). Η κλίση της γραμμής παλινδρόμησης ισοδυναμεί με τον κίνδυνο beta. Αυτό φαίνεται και από το διάγραμμα (10). Αν σχεδιάσουμε τη μορφή των αποδόσεων της αγοράς έναντι της απόδοσης της κάθε μετοχής και βρούμε την γραμμή παλινδρόμησης, τότε θα παρατηρήσουμε ότι η κλίση ισούται με την μονάδα, δηλαδή οι μετοχές μεταβάλλονται κατά το ίδιο ποσοστό με την αγορά. Παρόλα αυτά, αν η τιμή της μετοχής μεταβάλλεται λιγότερο από ότι μεταβάλλεται η αγορά, τότε η κλίση της γραμμής παλινδρόμησης θα είναι μικρότερη από την μονάδα, ενώ η κλίση θα είναι μεγαλύτερη από την μονάδα, όταν η τιμή της μετοχής μεταβάλλεται περισσότερο από ότι μεταβάλλεται η αγορά. Η αρνητική κλίση δείχνει ότι η μετοχή μετακινείται προς την αντίθετη κατεύθυνση από εκείνη της αγοράς. Για παράδειγμα αν η αγορά κινείται προς τα πάνω, η τιμή της μετοχής πέφτει προς τα κάτω.



ΔΙΑΓΡΑΜΜΑ (10)

Το φύλλο εργασίας Stock, που παίρνουμε από την βάση δεδομένων της επιχείρησης Tracway , περιλαμβάνει τις ημερήσιες τιμές της μετοχής Tracway για την περίοδο από 30 Ιουνίου μέχρι της 31 Δεκεμβρίου του 2000. Το διάγραμμα (11) είναι ένα διάγραμμα διασποράς της απόδοσης του δείκτη S&P500 και της απόδοσης της μετοχής Tracway για μια περίοδο έξι μηνών, αντίστοιχα. Φαίνεται καθαρά η συσχέτιση που εμφανίζεται να υπάρχει.



ΔΙΑΓΡΑΜΜΑ (11)

Η ποσοστιαία αύξηση (αρνητικές τιμές δείχνουν μείωση) και για τον S&P500 και για τις μετοχές της Tracway, υπάρχει στο φύλλο εργασίας Stock, από όπου παίρνουμε τα απαραίτητα στοιχεία για να δημιουργήσουμε το μοντέλο παλινδρόμησης.

Stock

Date	S&P500	Tracway	SAP % change	Tracway % change
30-Jun-00	15300	$ 100.00		
3-Jul-00	15438	$ 101.50	0.90%	1.50%
5-Jul-00	15867	$ 102.50	2.78%	0.99%
6-Jul-00	14984	$ 97.10	-5.57%	-5.27%
7-Jul-00	15468	$ 97.70	3.23%	0.62%
10-Jul-00	15608	$ 99.50	0.91%	1.84%
11-Jul-00	16218	$ 103.10	3.91%	3.62%

Ημερήσια μεταβολή στην τιμή της μετοχής Tracway = β_0 + β_1 μεταβολή του S&P500
Ο πίνακας (4) δείχνει τα αποτελέσματα της εφαρμογής της παλινδρόμησης από το Excel. Το μοντέλο που προκύπτει είναι :

Ημερήσια μεταβολή στην τιμή της μετοχής Tracway = 0,00124 + 0,62124 μεταβολή του S&P500
Η τιμή του συντελεστή προσδιορισμού (R^2 = 0,90) δείχνει ότι ένα μεγάλο ποσοστό της μεταβλητότητας εξηγείται από το μοντέλο. Η κλίση της γραμμής παλινδρόμησης, β_1 , (ο κίνδυνος beta της μετοχής Tracway) είναι 0,62. Αυτό δείχνει ότι η Tracway έχει μικρότερο απόδοση από την αγορά.

Stock

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0.9506849
R Τετράγωνο	0.9038017
Προσαρμοσμένο R Τετράγωνο	0.9030196
Τυπικό σφάλμα	0.010344
Μέγεθος δείγματος	125

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	1	0.1236491	0.1236491	1155.609	2.2E-64
Υπόλοιπο	123	0.0131609	0.000107		
Σύνολο	124	0.1368099			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95.0%*	*Υψηλότερο 95.0%*
Τεταγμένη επί την αρχή	0.0012396	0.0009261	1.338491	0.1832053	-0.0005936	0.0030728	-0.0005936	0.003073
S&P % change	0.62124	0.0182749	33.994249	2.2E-64	0.5850661	0.657414	0.5850661	0.657414

ΠΙΝΑΚΑΣ (4)

Πολλαπλή Γραμμική Παλινδρόμηση- ΕΦΑΡΜΟΓΗ

Employee Success

Duration	Reeducation	College gpa	Age	M/F	College Grad	Local
10	18	3.01	33	0	1	1
10	16	2.78	25	1	1	1
10	18	3.15	26	1	1	0
10	18	3.86	24	0	1	1
9.6	16	2.58	25	0	1	1
8.5	16	2.96	23	1	1	1
8.4	17	3.56	35	1	1	1

Το τμήμα των Ανθρώπινων Πόρων της επιχείρησης Tracway ξεχωρίζει για το μεγάλο ποσοστό του κύκλου εργασιών, που πραγματοποιεί στις πωλήσεις του προσωπικού της. Το τμήμα αυτό περιλαμβάνει κάποιες στρατηγικές πολιτικής που μπορούν να αναγνωρίσουν τα απαραίτητα χαρακτηριστικά που πρέπει να έχουν τα άτομα για να έχουν την δυνατότητα να παραμείνουν περισσότερο καιρό στην επιχείρηση. Παρόλα αυτά, σε μια πρόσφατη συνάντηση του προσωπικού, οι διευθύνοντες των ανθρώπινων πόρων δεν μπόρεσαν να συμφωνήσουν ποια πρέπει να είναι αυτά τα χαρακτηριστικά. Κάποιοι υποστήριζαν ότι τα χρόνια της εκπαίδευσης και ο βαθμός κολεγίου είναι οι σημαντικότεροι συντελεστές. Άλλοι υποστήριζαν ότι η πρόσληψη πιο ώριμων υποψηφίων θα οδηγούσε στην μεγαλύτερη παραμονή τους στην επιχείρηση. Για να μπορέσουν να καταλήξουν στο σωστό αποτέλεσμα, το προσωπικό συμφώνησε να πραγματοποιηθεί μια στατιστική μελέτη για να καθοριστεί η επίδραση που θα έχουν τα χρόνια εκπαίδευσης, ο βαθμός κολεγίου και η ηλικία που προσλαμβάνονται τα άτομα στην παραμονή τους στην επιχείρηση. Ένα δείγμα 40 πωλητών που προσβλήθηκαν δέκα χρόνια πριν ,επιλέχθηκε για να καθοριστεί η επίπτωση που είχαν αυτές οι μεταβλητές στην διάρκεια του κάθε άτομου που παρέμεινε στην επιχείρηση.

Στο παράδειγμα μας έχουμε μία εξαρτημένη μεταβλητή (τα χρόνια παραμονής στην επιχείρηση Tracway) και τρεις ανεξάρτητες μεταβλητές (τα χρόνια εκπαίδευσης, ο βαθμός κολεγίου, και η ηλικία). Το υπόδειγμα της απλής γραμμικής παλινδρόμησης που αναπτύξαμε προηγουμένως αναφέρεται σε σχέσεις που περιλαμβάνουν μία μόνο ερμηνευτική μεταβλητή. Η συμπεριφορά όμως των περισσότερων οικονομικών μεταβλητών είναι συνάρτηση όχι μιας αλλά πολλών μεταβλητών. Ένα μοντέλο παλινδρόμησης με περισσότερες από μία ανεξάρτητες μεταβλητές καλείται μοντέλο **πολυδιάστατης παλινδρόμησης**. Αν όλοι οι όροι στο μοντέλο είναι γραμμικοί, τότε έχουμε το μοντέλο της **γραμμικής πολυμεταβλητής παλινδρομήσεως**. Η απλή γραμμική παλινδρόμηση είναι μια ειδική περίπτωση της πολλαπλής γραμμικής παλινδρόμησης.

Ένα μοντέλο πολυδιάστατης γραμμικής παλινδρόμησης έχει την εξής μορφή:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$$

,όπου το Y: είναι η εξαρτημένη μεταβλητή
$X_{1...}X_k$: είναι οι ανεξάρτητες ή αλλιώς ερμηνευτικές μεταβλητές
β_0 :είναι ο ερμηνευτικός όρος
β_1 :είναι οι συντελεστές παλινδρόμησης για τις ανεξάρτητες μεταβλητές
e :είναι το τυπικό σφάλμα

Οι βασικές υποθέσεις που συνιστούν το κλασικό γραμμικό υπόδειγμα στη γενική του μορφή, δηλαδή με Κ ερμηνευτικές μεταβλητές, είναι σχεδόν οι ίδιες με τις υποθέσεις για το διμεταβλητό γραμμικό υπόδειγμα. Οι υποθέσεις αυτές, που πρέπει να ισχύουν για όλες τις παρατηρήσεις, είναι οι ακόλουθες :

1. $Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + ... + \beta_k X_{tk} + e_t$
2. $e_t \sim (0, \sigma^2)$

   α) e_t *είναι τυχαία μεταβλητή*
   β) $Ee_t = 0$
   γ) $Ee_t^2 = \sigma^2$

3. $Ee_t e_s = 0$ *για* $t \neq s$

4. *Η μεταβλητή X δεν είναι στοχαστική. Οι τιμές της παραμένουν σταθερές και δεν είναι όλες ίσες μεταξύ τους.*

5. *Δεν υπάρχουν ακριβείς γραμμικές σχέσεις ανάμεσα στις ερμηνευτικές μεταβλητές.*

6. *Ο αριθμός των παρατηρήσεων του δείγματος είναι μεγαλύτερος από τον αριθμό των συντελεστών του υποδείγματος που θέλουμε να εκτιμήσουμε.*

Η υπόθεση 1 αναφέρεται στην γραμμική σχέση που συνδέει τις μεταβλητές Y και X_1, X_2, ..., X_k. Κάθε τιμή t της εξαρτημένες μεταβλητής είναι γραμμική συνάρτηση των τιμών των ερμηνευτικών μεταβλητών X_{t1}, X_{t2}, ..., X_{tk} και του διαταρακτικού όρου e_t. Οι υποθέσεις 2 και 3 αναφέρονται στον διαταρακτικό όρο e_t και είναι ακριβώς ίδιες με τις αντίστοιχες υποθέσεις του απλού υποδείγματος. Η υπόθεση 4 είναι επίσης ακριβώς ίδια με την αντίστοιχη υπόθεση του γραμμικού υποδείγματος, με την διαφορά ότι τώρα αναφέρεται σε Κ ερμηνευτικές μεταβλητές. Οι πρόσθετες υποθέσεις 5 και 6 έχουν σχέση με την εκτίμηση και τον έλεγχο του υποδείγματος. Η υπόθεση 5 αποτελεί προϋπόθεση για την εκτίμηση του υποδείγματος και αποκλείει την ύπαρξη πολυσυγγραμμικότητας μεταξύ των ερμηνευτικών μεταβλητών. Αυτό σημαίνει πως καμία από τις Κ ερμηνευτικές δεν μπορεί να εκφραστεί ως γραμμικός συνδυασμός των υπολοίπων. Τέλος, η υπόθεση 6 εξασφαλίζει τους απαραίτητους βαθμούς ελευθερίας και για την εκτίμηση αλλά και για τον έλεγχο του υποδείγματος. Ο αριθμός των παρατηρήσεων πρέπει να είναι τουλάχιστον ίσος με τους συντελεστές του υποδείγματος, για να είναι δυνατή η εκτίμηση του, όπως θα φανεί αργότερα. Πρέπει όμως να είναι και μεγαλύτερος, για να είναι δυνατός ο έλεγχος υποθέσεων με

τις διάφορες στατιστικές ελέγχου που η κατανομή τους εξαρτάται από τους βαθμούς ελευθερίες, όπως η κατανομή t ή η κατανομή F.

Για την επιχείρηση Tracway όπως είπαμε έχουμε τρεις ανεξάρτητες μεταβλητές και έτσι το μοντέλο γράφεται ως εξής :

Χρόνια παραμονής στην επιχείρηση (retention) = β_0 + β_1Χρόνια εκπαίδευσης(Years education) + β_2 Βαθμός κολεγίου (GPA) + β_3 Ηλικία (Age) + e

Όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης, εκτιμούμε τους συντελεστές παλινδρόμησης, που καλούνται μερικοί συντελεστές παλινδρόμησης (partial regression coefficients), όπως το b_0 ,b_1 ….b_k και έπειτα μπορούμε να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής χρησιμοποιώντας το υπόδειγμα :

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_k X_k$$

Οι συντελεστές μερικής παλινδρόμησης παριστάνουν την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής, όταν η αντίστοιχη εξαρτημένη μεταβλητή αυξάνεται κατά μία μονάδα ενώ οι τιμές των άλλων ανεξάρτητων μεταβλητών παραμένουν σταθερές. Επιπλέον, το β_2 αναπαριστάνει την μεταβολή στην παραμονή στην επιχείρηση όταν το GPA αυξάνεται κατά μία μονάδα, ενώ τα χρόνια εκπαίδευσης και η ηλικία παραμένουν σταθερά.

Όπως και στην απλή γραμμική παλινδρόμηση, η πολλαπλή γραμμική παλινδρόμηση χρησιμοποιεί την αρχή των ελαχίστων τετραγώνων στην εκτίμηση του ερμηνευτικού όρου και του συντελεστή κλίσης που ελαχιστοποιεί τους τετραγωνισμένους όρους των σφαλμάτων όλων των παρατηρήσεων. Αυτός είναι ένας από τους λόγους που οι βασικές υποθέσεις που πρέπει να ισχύουν στην απλή γραμμική παλινδρόμηση, πρέπει να ισχύουν και στην πολλαπλή.

Γνωρίζουμε τα αποτελέσματα των υπαλλήλων της επιχείρησης Tracway. Χρησιμοποιώντας την ανάλυση δεδομένων βρίσκουμε την πολλαπλή παλινδρόμηση των στοιχείων αυτών, πίνακας (5). Η μόνη διαφορά με την απλή γραμμική παλινδρόμηση είναι ότι τα στοιχεία για τις ανεξάρτητες μεταβλητές βρίσκονται σε τρεις στήλες. Ο έλεγχος των καταλοίπων επιβεβαιώνει τις υποθέσεις που πρέπει να ισχύουν.

Employee Success

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0.388
R Τετράγωνο	0.15
Προσαρμοσμένο R Τετράγωνο	0.079
Τυπικό σφάλμα	2.726
Μέγεθος δείγματος	40

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	3	47.27	15.8	2.12101	0.115
Υπόλοιπο	36	267.4	7.43		
Σύνολο	39	314.7			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95.0%*	*Υψηλότερο 95.0%*
Τεταγμένη επί την αρχή	-2.74	4.504	-0.61	0.54721	-11.9	6.397718	-11.9	6.397718
Yrseducation	-0.07	0.355	-0.19	0.85131	-0.79	0.653252	-0.79	0.653252
College gpa	0.68	1.184	0.57	0.56918	-1.72	3.080332	-1.72	3.080332
Age	0.292	0.135	2.16	0.03761	0.018	0.565417	0.02	0.565417

ΠΙΝΑΚΑΣ (5)

Από τον πίνακα (5) φαίνεται ότι το μοντέλο έχει την μορφή :

Retention = -2,74-0,067 Years education + 0,680 GPA + 0,292 Age

Για παράδειγμα , για ένα άτομο ηλικίας 30 ετών με 15 χρόνια εκπαίδευσης και με βαθμό κολεγίου 2.50 , το μοντέλο παλινδρόμησης μπορεί να προβλέψει ότι αυτός ο υποψήφιος θα παραμείνει στην εταιρεία :

Retention = -2,74 - 0,067 × 16 + 0,060 × 2,50 + 0,292 × 30 = 6,648 years

Ένα σημείο που πρέπει να τονίσουμε είναι ότι είναι επικίνδυνο να παρατείνεις ένα μοντέλο παλινδρόμησης έξω από τις σειρές που καλύπτονται από τις παρατηρήσεις. Για παράδειγμα για ένα άτομο ηλικίας 40 ετών με 13 χρόνια εκπαίδευσης και με μέσο όρο βαθμό κολεγίου 1,5, η παλινδρόμηση θα προβλέψει 9,1 χρόνια παραμονής στην επιχείρηση. Αυτό μπορεί να μην ανταποκρίνεται στην πραγματικότητα, αλλά στο μοντέλο παλινδρόμησης δεν μελετούνται στοιχεία με αυτά τα χαρακτηριστικά.

Ερμηνεία αποτελεσμάτων της Πολλαπλής Γραμμικής Παλινδρόμησης

Τα αποτελέσματα της ανάλυσης της παλινδρόμησης είναι της ίδιας μορφής με αυτά της γραμμικής παλινδρόμησης. Το πολλαπλό R, ο πολλαπλός συντελεστής συσχέτισης ,και το R τετράγωνο ,ο πολλαπλός συντελεστής προσδιορισμού, δείχνουν την δύναμη της σχέσης μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών. Η χαμηλή τιμή του R^2, R^2 = 0,15, δείχνει ότι μόνο 15% της μεταβλητότητας των στοιχείων εξηγείται από τις ανεξάρτητες μεταβλητές, ένα προφανώς φτωχό συνταίριασμα..

1). **Είναι το μοντέλο στατιστικά σημαντικό ;** .

Η ANOVA στον πίνακα (5) ελέγχει την σημαντικότητα όλου του μοντέλου. Δηλαδή υπολογίζει την στατιστική F για να πραγματοποιηθούν οι έλεγχοι των υποθέσεων.

H_0: $\beta_1 = \beta_2 = \beta_k = 0$
H_1: τουλάχιστον ένα από τα β είναι διάφορο του μηδενός

Η μηδενική υπόθεση δηλώνει ότι δεν υπάρχει γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και κάθε μία από τις ανεξάρτητες μεταβλητές, ενώ η εναλλακτική υπόθεση δηλώνει ότι η εξαρτημένη μεταβλητή παρουσιάζει γραμμική σχέση με μία τουλάχιστον από τις ανεξάρτητες μεταβλητές. Δεν μπορούμε να καταλήξουμε στην ύπαρξη γραμμικής σχέσης της εξαρτημένης μεταβλητής με όλες τις ανεξάρτητες μεταβλητές. Ο έλεγχος είναι πανομοιότυπος με αυτόν της απλής γραμμικής παλινδρόμησης. Η στατιστική F υπολογίζεται από το κλάσμα MSR / MSE, εκτός αν αυτό έχει k και n-k-1 βαθμούς ελευθερίας. Σε ένα επίπεδο σημαντικότητας 5%, θα αποτύχουμε να απορρίψουμε την μηδενική υπόθεση γιατί η *σημαντικότητα του F* είναι μεγαλύτερη από 0,05. Μια άλλη εξήγηση είναι ότι η πιθανότητα να πάρουμε αυτά τα αποτελέσματα από ένα τυχαίο δείγμα ενός πληθυσμό, στο οποίο δεν υπάρχει καμία σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών, είναι 0,114, δηλαδή το επίπεδο σημαντικότητας με ποσοστό 5% βρίσκεται αρκετά πιο ψηλά.

2). **Είναι οι μεταβλητές από μόνες τους σημαντικές;**

2. Το τελευταίο μέρος στα αποτελέσματα, προβάλλει τις πληροφορίες για τον έλεγχο των υποθέσεων για τον κάθε ένα συντελεστή παλινδρόμησης ξεχωριστά. Για παράδειγμα, ο έλεγχος της υπόθεσης για την κλίση του πληθυσμού β_1(που αντιστοιχεί στα χρόνια εκπαίδευσης) είναι μηδέν, υπολογίζουμε την στατιστική t διαιρώντας τον εκτιμητή b_1
(-0,06705) με το τυπικό του σφάλμα (0,355165) και βρίσκουμε ότι ισούται με –0,1888, όπως βλέπουμε και στην τρίτη στήλη. Αυτή η στατιστική έχει n-k-1 βαθμούς ελευθερίας , ή σε αυτήν την περίπτωση , 40-30-1 = 36. Γνωρίζουμε από τους στατιστικούς πίνακες κατανομών ότι $t_{36,0.025} = 2,0281$. Έτσι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 0,05. Αυτό το αποτέλεσμα μπορεί επίσης να προκύψει εξετάζοντας την τιμή-P (P-value) στην επόμενη στήλη των αποτελεσμάτων.

Από την άλλη πλευρά, με τον έλεγχο t μπορούμε να συμπεράνουμε ότι η κλίση που αντιστοιχεί στην ηλικία είναι σημαντική παρόλο που δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση , όσον αφορά την σημαντικότητα όλου του μοντέλου. Τέτοιες ανωμαλίες μπορούν να συμβούν σε μοντέλα πολλαπλής παλινδρόμησης για πολλούς λόγους, περιλαμβάνοντας την αλληλεξάρτηση μεταξύ των μεταβλητών. Ο έλεγχος F δείχνει την σημαντικότητα όλου του μοντέλου. Στην περίπτωσή μας, η χαμηλή τιμή R^2 υποδηλώνει ένα φτωχό συνταίριασμα, παρόλο που η ηλικία από μόνη της εμφανίζει σημαντικότητα., σε συνδυασμό με τις άλλες μεταβλητές δεν μπορούμε να καταλήξουμε στο συμπέρασμα ότι όλες οι κλίσεις είναι διαφορετικές του μηδενός.

Μπορούμε επίσης να χρησιμοποιήσουμε τις αποδόσεις της παλινδρόμησης για να υπολογίσουμε τα διαστήματα εμπιστοσύνης για την κλίση των συντελεστών. Ένα διάστημα εμπιστοσύνης για το β_j θα μπορούσε να είναι:

$$\beta_j \pm t_{n-k-1}\, s.e$$

όπου s.e είναι το τυπικό σφάλμα που φαίνεται στην συνοπτική απόδοση.

3). **Υπάρχει Πολυσυγραμμικότητα;**

……..

Η **συσχέτιση** (correlation) είναι μια αριθμητική τιμή μεταξύ του –1 και 1, και μετρά την γραμμική σχέση μεταξύ των μεταβλητών. Όσο υψηλότερη είναι η απόλυτη τιμή της συσχέτισης , τόσο περισσότερο δυνατή είναι η σχέση μεταξύ των μεταβλητών. Το πρόσημο δείχνει εάν η σχέση μεταξύ των μεταβλητών είναι θετική ή αρνητική. Μέσω Excel ,με το εργαλείο συσχέτισης , μπορούμε να υπολογίσουμε την συσχέτιση μεταξύ των συνδυασμένων μεταβλητών. Ο πίνακας (7) δείχνει τον πίνακα συσχέτισης για τα διατομικά στοιχεία που συλλέχθηκαν κατά την διάρκεια της παραμονής των εργαζομένων στην επιχείρηση Tracway. Μπορούμε να δούμε ότι η ηλικία έχει την δυνατότερη σχέση με την εξαρτημένη μεταβλητή, από ότι τα χρόνια εκπαίδευσης και ο βαθμός του κολεγίου. Αυτός ο πίνακας δίνει πληροφορίες για το ποια ανεξάρτητη μεταβλητή θα επιλεχθεί πρώτα για τον συνυπολογισμό της στο μοντέλο.

	Duration	Yrseducation	*College gpa*	*Age*
Duration	1			
Yrseducation	0.1796535	1		
College gpa	0.17743653	0.58665385	1	
Age	0.3766582	0.42102889	0.248521587	1

ΠΙΝΑΚΑΣ (7)

Η **πολυσυγγραμμικότητα** (multicollinearity) είναι χαρακτηριστικό του δείγματος, αναφέρεται δηλαδή στις γραμμικές σχέσεις ανάμεσα στις ερμηνευτικές μεταβλητές στο δείγμα και όχι στον πληθυσμό. Αυτό σημαίνει πως δεν μπορεί να γίνει έλεγχος, με την στατιστική έννοια του όρου, για πολυσυγγραμμικότητα. Οι διάφοροι τρόποι « ελέγχου » της πολυσυγγραμμικότητας αναφέρονται στην διαπίστωση και μέτρηση και όχι στον έλεγχο της πολυσυγγραμμικότητας.

Όταν το υπόδειγμα περιλαμβάνει δύο μόνο ερμηνευτικές μεταβλητές, η τιμή του συντελεστή απλής συσχετίσεως μεταξύ των ερμηνευτικών μεταβλητών είναι ικανοποιητικό μέτρο του βαθμού της πολυσυγγραμμικότητας που υπάρχει στο δείγμα. Όσο όμως αυξάνει ο αριθμός των ερμηνευτικών μεταβλητών τόσο δυσκολότερη γίνεται η διαπίστωση και η μέτρηση της πολυσυγγραμμικότητας. Η εξέταση των συντελεστών απλής συσχετίσεως ανάμεσα στις ερμηνευτικές μεταβλητές δεν είναι αρκετή, γιατί όπως αναφέραμε παραπάνω, μπορεί η τιμή των συντελεστών συσχετίσεως να είναι χαμηλή και το δείγμα να χαρακτηρίζεται από πολυσυγγραμμικότητα.

Συμπεριλαμβάνοντας πολλές μεταβλητές στο μοντέλο παλινδρόμησης ,μπορεί να οδηγηθούμε σε πολυσυγγραμμικότητα. Είναι μια κατάσταση στην οποία δύο ή

περισσότερες μεταβλητές στο ίδιο μοντέλο παλινδρόμησης έχουν υψηλά επίπεδα της ίδιας πληροφορίας, εξηγώντας την μεταβλητότητα της εξαρτημένης μεταβλητής και την συσχέτιση της μίας με την άλλη. Συγκεκριμένα, η πολυσυγγραμμικότητα υπάρχει, όταν σε ένα σύνολο από ανεξάρτητες μεταβλητές, η μία ανεξάρτητη μεταβλητή προβλέπει την άλλη καλύτερα από ότι προβλέπουν την εξαρτημένη μεταβλητή. Η πολυσυγγραμμικότητα καθιστά δύσκολη την διάκριση των επιρροών που έχουν αυτές οι μεταβλητές στην εξαρτημένη μεταβλητή.

Το πρόβλημα της παλινδρόμησης, δηλαδή ,υφίσταται όταν δεν έχουμε σωστές φυσικές ερμηνείες στα αποτελέσματα της ανάλυσης της παλινδρόμησης. Για παράδειγμα θα είχαμε το πρόβλημα της πολυσυγγραμμικότητας αν για την ανεξάρτητη μεταβλητή, την ηλικία, παρατηρούσαμε ότι καθώς αυξανόταν η ηλικία αυξανόταν και ο συντελεστής αυτής της ανεξάρτητης μεταβλητής, ενώ στην πραγματικότητα θα έπρεπε να μειώνεται. Ένας άλλος τρόπος για να δούμε αν υπάρχει το πρόβλημα της πολυσυγγραμμικότητας είναι να βρούμε τις μεταξύ συσχετίσεις των μεταβλητών. Αν είναι θετικές τότε έχουμε το πρόβλημα της πολυσυγγραμμικότητας. Επιπλέον, ένας άλλος τρόπος εύρεσης της ύπαρξης πολυσυγγραμμικότητας είναι μέσω του παράγοντα επιρροής διακύμανσης (variance inflation factor)

Η πολυσυγγραμμικότητα μπορεί να συντελέσει έμμεσα σε λανθασμένη εξειδίκευση του υποδείγματος. Αυτό μπορεί να συμβεί, γιατί στην πράξη , πολλές φορές προσθέτουμε ή αφαιρούμε ερμηνευτικές μεταβλητές ανάλογα με το αν είναι στατιστικά σημαντικές ή όχι. Αν όμως η μη σημαντικότητα του συντελεστή οφείλεται στην ύπαρξη πολυσυγγραμμικότητας, η αφαίρεση της σχετικής μεταβλητής δημιουργεί σφάλματα εξειδικεύσεως.

Από την προηγούμενη ανάλυση μπορούμε να συμπεράνουμε, πως η πολυσυγγραμμικότητα είναι σοβαρό πρόβλημα, γιατί επηρεάζει την αξιοπιστία των αποτελεσμάτων της εκτιμήσεως. Οι πιο σοβαρές από τις συνέπειές της αναφέρονται : α) στην ακρίβεια των συντελεστών, επειδή οι διακυμάνσεις μπορεί να είναι σχετικά μεγάλες, β) στη σταθερότητα των συντελεστών, και γ) στη δυνατότητα σφάλματος εξειδίκευσης. Πρέπει όμως να τονισθεί ότι η πολυσυγγραμμικότητα δεν επηρεάζει τις ιδιότητες των εκτιμητών που παίρνουμε με την μέθοδο των ελαχίστων τετραγώνων. Οι εκτιμητές δηλαδή, εξακολουθούν να είναι άριστοι γραμμικοί αμερόληπτοι.

Όταν υπάρχει πολυσυγγραμμικότητα, οι συντελεστές β των ανεξάρτητων μεταβλητών είναι ευμετάβλητοι, και ακόμα και το πρόσημό τους είναι πιθανόν να αλλάζει όταν συμπεριλαμβάνονται διαφορετικές μεταβλητές. Επίσης οι τιμές P μπορεί να διογκωθούν, και αυτό θα έχει σαν αποτέλεσμα να μην απορριφθεί η μηδενική υπόθεση, ενώ θα έπρεπε να απορριφθεί στην πραγματοποίηση του ελέγχου για την σημαντικότητα του μοντέλου. Η αναφορά στην πολυσυγγραμμικότητα έγινε για να πληροφορηθούμε για τα προβλήματα που είναι πιθανόν να προκαλέσει και έτσι σε μια τέτοια περίπτωση θα είναι απαραίτητη η συμβουλή ενός ειδικού.

4). Ανάλυση καταλοίπων (Επιβεβαίωση παραδοχών)

…….

Όπως και σε κάθε μοντέλο παλινδρόμησης, χρειάζεται να παρατηρούμε προσεκτικά τα αποτελέσματα, για να μπορέσουμε να τα κατανοήσουμε. Τα αποτελέσματα δεν υπονοούν ότι περισσότερη εκπαίδευση δεν θα είναι χρήσιμη, ή ότι ο υψηλότερος μέσος όρος μαθημάτων κολεγίου δεν μπορεί να συμβάλει στην επιτυχία. Θέλει να πει ότι εκείνοι με περισσότερη μόρφωση παρουσιάζουν ελαστικότητα εργασίας, δηλαδή μπορούν εύκολα να μετακινούνται σε άλλες εργασίες. Το ίδιο συμβαίνει και με τον μέσο όρο του βαθμού απολυτηρίου. Αλλά όταν συγκεκριμένα μελετάμε πόσο καιρό, τα σαράντα άτομα του δείγματος που προσβλήθηκαν, παρέμειναν στην επιχείρηση μετά από δέκα χρόνια, εκείνοι που ήταν μεγαλύτεροι παρέμειναν περισσότερο στην επιχείρηση Tracway. Περισσότερα χρόνια εκπαίδευσης καθώς και υψηλότερος μέσος όρος μαθημάτων δεν ήταν σημαντικοί παράγοντες για να παραμείνουν περισσότερο χρονικό διάστημα στην επιχείρηση. Άλλες μεταβλητές που δεν συμπεριλαμβάνονται στο μοντέλο, είναι πιθανόν να οδηγούσαν σε διαφορετικά αποτελέσματα. Έτσι προκύπτει η απορία, πώς να δημιουργήσουμε « καλά » μοντέλα παλινδρόμησης.

Δημιουργία « καλών » μοντέλων Παλινδρόμησης

Ένα καλό μοντέλο παλινδρόμησης θα πρέπει να περιλαμβάνει μόνο στατιστικά σημαντικές ανεξάρτητες μεταβλητές. Επειδή τα χρόνια εκπαίδευσης και ο μέσος όρος βαθμός κολεγίου δεν φαίνεται να επηρεάζουν σημαντικά την ανεξάρτητη μεταβλητή, πρέπει να τις παραλείψουμε, και να υπολογίσουμε το μοντέλο, περιλαμβάνοντας μόνο την ηλικία. Τα αποτελέσματα αυτού φαίνονται στον πίνακα (6). Το καινούργιο μοντέλο είναι :

$$\text{Retention} = -2{,}01 + 0{,}300 \times \text{Age}$$

Σε αυτό το μοντέλο παλινδρόμησης το R^2 είναι 0,142 αντί 0,150 που είναι στο μοντέλο και με τις τρεις ανεξάρτητες μεταβλητές, δηλαδή το μοντέλο και με τις τρεις ανεξάρτητες μεταβλητές είναι λίγο καλύτερο από το καινούργιο μοντέλο . Η ηλικία είναι στατιστικά σημαντική μεταβλητή, με σημαντικότητα F ίση με 0,017. Αυτό δείχνει ότι υπάρχει ακόμα μεγαλύτερη σημαντικότητα από ότι στο προηγούμενο μοντέλο. Η πρόβλεψη για τους τριαντάχρονους υποψηφίους με 16 χρόνια εκπαίδευσης και μέσο όρο βαθμό κολεγίου 2,50, είναι ότι θα παραμείνουν στην επιχείρηση 6,99 χρόνια, λίγο περισσότερο από ότι στο προηγούμενο μοντέλο, γιατί οι άλλες δύο μεταβλητές δεν συμπεριλαμβάνονται στο καινούργιο μοντέλο.

Employee Success

Στατιστικά παλινδρόμησης	
Πολλαπλό R	0.3766582
R Τετράγωνο	0.1418714
Προσαρμοσμένο R Τετράγωνο	0.11928907
Τυπικό σφάλμα	2.66580544
Μέγεθος δείγματος	40

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	1	44.64604247	44.64604	6.282407	0.01659
Υπόλοιπο	38	270.0477075	7.106519		
Σύνολο	39	314.69375			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατώτερο 95%*	*Υψηλότερο 95%*	*Κατώτερο 95.0%*	*Υψηλότερο 95.0%*
Τεταγμένη επί την αρχή	-2.01486568	3.042483099	-0.66224	0.511812	-8.17405	4.1443196	-8.174051	4.14432
Age	0.30029287	0.119806943	2.506473	0.016592	0.05776	0.5428294	0.057756	0.54283

ΠΙΝΑΚΑΣ (6)

Α) Χρήση διορθωμένου συντελεστή προσδιορισμού (Adjusted R^2) στον υπολογισμό του ταιριάσματος

Ένας χρήσιμος τρόπος για να εξετάσουμε το σχετικό ταίριασμα των διαφορετικών μοντέλων, είναι μέσω του διορθωμένου συντελεστή προσδιορισμού. Προσθέτοντας μια ανεξάρτητη μεταβλητή στο μοντέλο παλινδρόμησης, ο συντελεστής προσδιορισμού που θα προκύπτει θα είναι ίσος η μεγαλύτερος από αυτόν του πρωταρχικού μας μοντέλου. Αυτό ισχύει ακόμα και όταν η νέα ανεξάρτητη μεταβλητή έχει κάποια σχέση με την εξαρτημένη μεταβλητή. Ο διορθωμένος συντελεστής προσδιορισμού απεικονίζει τον αριθμό των ανεξάρτητων μεταβλητών που υπάρχουν στο μοντέλο αλλά και το μέγεθος του δείγματος. Αυτό βοηθάει στο να κατανοήσουμε καλύτερα την τιμή της ανεξάρτητης μεταβλητής που προστέθηκε στο μοντέλο. Αν εκφράσουμε το συνολικό άθροισμα των τετραγώνων και το άθροισμα των τετραγώνων των καταλοίπων ως διακυμάνσεις, δηλαδή αν τα διαιρέσουμε με τους αντίστοιχους βαθμούς ελευθερίας, ώστε να έχουμε αμερόληπτους εκτιμητές των αντίστοιχων διακυμάνσεων στον πληθυσμό, παίρνουμε τον **διορθωμένο συντελεστή προσδιορισμού**. Ο διορθωμένος συντελεστής προσδιορισμού, δηλαδή , στην πολλαπλή γραμμική παλινδρόμηση δίνεται από την σχέση:

$$\text{Adjusted } R^2 = 1 - \frac{SSE}{SST}\left(\frac{n-1}{n-k-1}\right)$$

,όπου το SSE: είναι άθροισμα των τετραγώνων των καταλοίπων
SST: είναι το συνολικό άθροισμα τετραγώνων
n: είναι ο αριθμός των παρατηρήσεων
k: είναι ο αριθμός των ανεξάρτητων μεταβλητών

Συγκεκριμένα ο προσαρμοσμένος όρος (n-1) / (n-k-1) παριστάνει την αναλογία των βαθμών ελευθερίας του SSE και του SST. Ο διορθωμένος συντελεστής προσδιορισμού είναι περισσότερο κατάλληλος για την σύγκριση της ερμηνευτικής ικανότητας των υποδειγμάτων, όταν ο αριθμός των ερμηνευτικών μεταβλητών καθώς και το μέγεθος του δείγματος διαφέρουν. Από τη σχέση του διορθωμένου συντελεστή προσδιορισμού είναι φανερό ότι αυτός είναι μικρότερος από τον συντελεστή προσδιορισμού. Θα είναι ίσοι μόνο όταν $R^2 = 1$ ή ασυμπτωτικά, δηλαδή όταν το μέγεθος του δείγματος τείνει στο άπειρο. Επιπλέον ο συντελεστής προσδιορισμού δεν μπορεί να πάρει αρνητικές τιμές ενώ ο διορθωμένος συντελεστής προσδιορισμού μπορεί.

Για την επιχείρηση Tracway, στο μοντέλο και με τις τρεις ανεξάρτητες μεταβλητές, βρίσκουμε ότι:

$$\text{Adjusted } R^2 = 1 - \frac{267{,}4259}{314{,}6938}\left(\frac{39}{36}\right) = 0{,}0794$$

Για το μοντέλο παλινδρόμησης με μία μόνο ανεξάρτητη μεταβλητή, την ηλικία, ο διορθωμένος συντελεστής προσδιορισμού είναι 0,119. Παρατηρείται ότι ενώ το R^2 θα έπρεπε να μειώνεται παραλείποντας τις άλλες δύο μεταβλητές από το μοντέλο, στην πραγματικότητα αυξάνεται, δείχνοντας ένα καλύτερο μοντέλο ταιριάσματος. Επιπλέον, ο διορθωμένος συντελεστής προσδιορισμού μας δίνει την δυνατότητα να υπολογίζουμε την επίδραση της πρόσθεσης ή της μετακίνησης μεταβλητών μέσα ή έξω από το μοντέλο.