Στο αρχείο δεδομένων dummy1.sav καταγράφονται τα χρόνια εμπειρίας (exprnc), το επίπεδο μόρφωσης (educ), οι αρμοδιότητες (mgt) και ο μισθός (salary) 46 υπαλλήλων. Να βρεθεί ένα μοντέλο πρόβλεψης του μισθού με βάση τα πιο πάνω δεδομένα.

Σχόλια για το αρχείο dummy1.sav

1. Η μόρφωση είναι κατηγορική μεταβλητή με τρεις τιμές οπότε μετατρέπεται σε δύο δείκτριες μεταβλητές τις educ1 & educ2.
2. Προσαρμόζω το μοντέλο υποθέτοντας ότι the effect of education and management status on salary determination are additive:

A linear relationship will be used for salary and experience. We shall assume that each additional year of experience is worth a fixed salary increment. Education may also be treated in a linear fashion. If the education variable is used in the regression equation in raw form, we would be assuming that each step up in education is worth a fixed increment to salary. That is, with all other variables held constant, the relationship between salary and education is linear. That interpretation is possible but may be too restrictive. Instead, we shall view education as a categorical variable and define two indicator variables to represent the three categories. These two variables allow us to pick up the effect of education on salary whether or not it is linear. The management variable is also an indicator variable designating the two categories, 1 for management positions and 0 for regular staff positions it follows

that there is a different regression equation for each of the six (three education and two management) categories as shown in Table 5.2. According to the proposed model, we may say that the indicator variables help to determine the base salary level as a function of education and management status after adjustment for years of experience.

Regression Equations for the Six Categories

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	11031,808	383,217		28,787	,000
	exprnc	546,184	30,519	,599	17,896	,000
	High school	-2996,210	411,753	-,296	-7,277	,000
	College	147,825	387,659	,016	,381	,705
	mgt	6883,531	313,919	,731	21,928	,000

a Dependent Variable: salary

That is, each additional year of experience is estimated to be worth an annual salary increment of $546. The other coefficients may be interpreted by looking into Table 5.2. The coefficient of the management indicator variable, 61, is estimated to be 6883.50. From Table

5.2 we interpret this amount to be the average incremental value in annual salary associated with a management position. For the education variables, `y1` measures

Σχόλιο [a1]: 6 μοντέλα Γιατί 2 επίπεδα της mgt 3 της educ.

Σχόλιο [a2]: Μετρά τη διαφορά στο μισθό μεταξύ HS και Advanced. Άρα ένας της κατηγορίας Advanced κερδίζει 2996 περισσότερα. Ένας του κολλεγίου κερδίζει 147 δολάρια περισσότερα αλλά αυτή η διαφορά δεν είναι στατιστικά σημαντική. Ένας του κολεγίου κερδίζει 2996+147=3144 δολάρια περισσότερα.

Σχόλιο [a3]: Μη στατιστικά σημαντικό

Σχόλιο [a4]: Ερμηνεία? π.χ αύξηση εμπειρίας κατά ένα έτος αύξηση μισθού κατά 546 δολάρια

the salary differential for the H.S. category relative to the advanced degree category and γ_2 measures the differential for the **B.S.** category relative to the advanced degree category. The difference, $\gamma_2 - \gamma_1$, measures the differential salary for the H.S. category relative to the **B.S.** category. From the regression results, in terms of salary for computer professionals, we see that an advanced degree is worth $2996 more than a high school diploma, a **B.S.** is worth $148 more than an advanced degree (this differential is not statistically significant, $t = 0.38$), and a **B.S.** is worth about $3144 more than a high school diploma. These salary differentials hold for every fixed level of experience.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,978(a)	,957	,953	1027,43725

a Predictors: (Constant), mgt, exprnc, High school, College

Σχόλιο [a5]: ικανοποιητικό

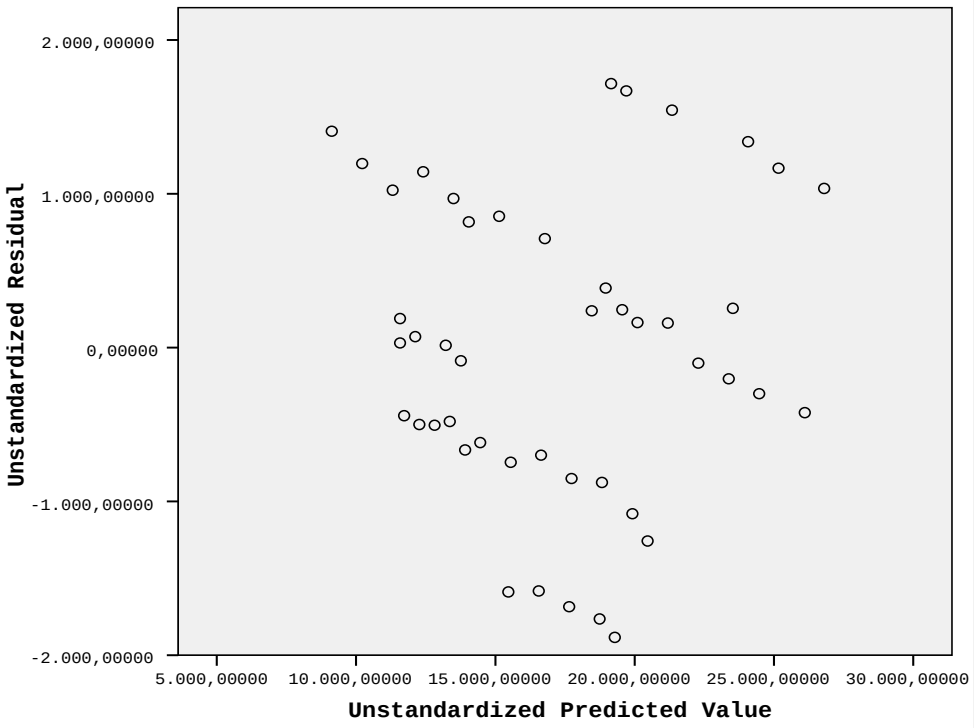
Ας εξετάσουμε κάποιες από τις υποθέσεις του μοντέλου

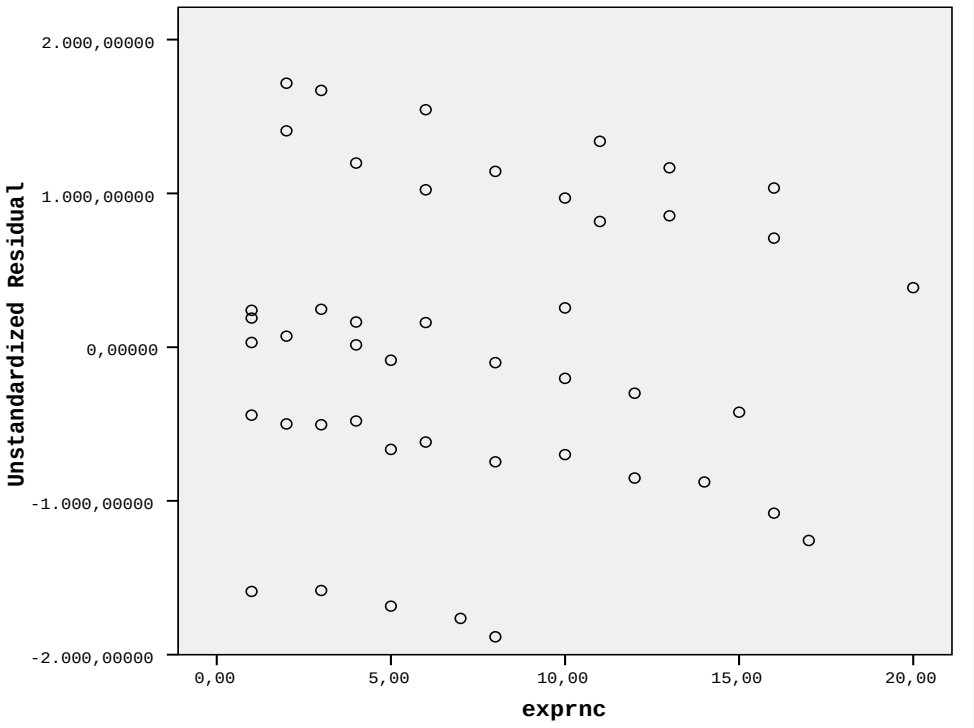
Π.χ ορθότητα μοντέλου όπως θα δούμε αναλυτικότερα σε άλλη ενότητα

Γραφική παράσταση των υπολοίπων ως προς τις ανεξάρτητες μεταβλητές. Αν δεν παρατηρηθεί κάποια ιδιαίτερη μορφή και τα σημεία βρίσκονται τυχαία γύρω από το μηδέν το μοντέλο μπορεί να θεωρηθεί ορθό. Αν δούμε κάποια ιδιαίτερη γραφική παράσταση τότε η εξαρτημένη και η ανεξάρτητη μεταβλητή μπορεί να μην συνδέονται με μία γραμμική σχέση.

Αποθηκεύω υπόλοιπα και εκτιμώμενες τιμές.

Graphs Chart Builder Scatter Dot Simple


Unstandardized Residual
2.000,00000
1.000,00000
0,00000
-1.000,00000
-2.000,00000
5.000,00000
10.000,00000
15.000,00000
20.000,00000
25.000,00000
30.000,00000
Unstandardized Predicted Value

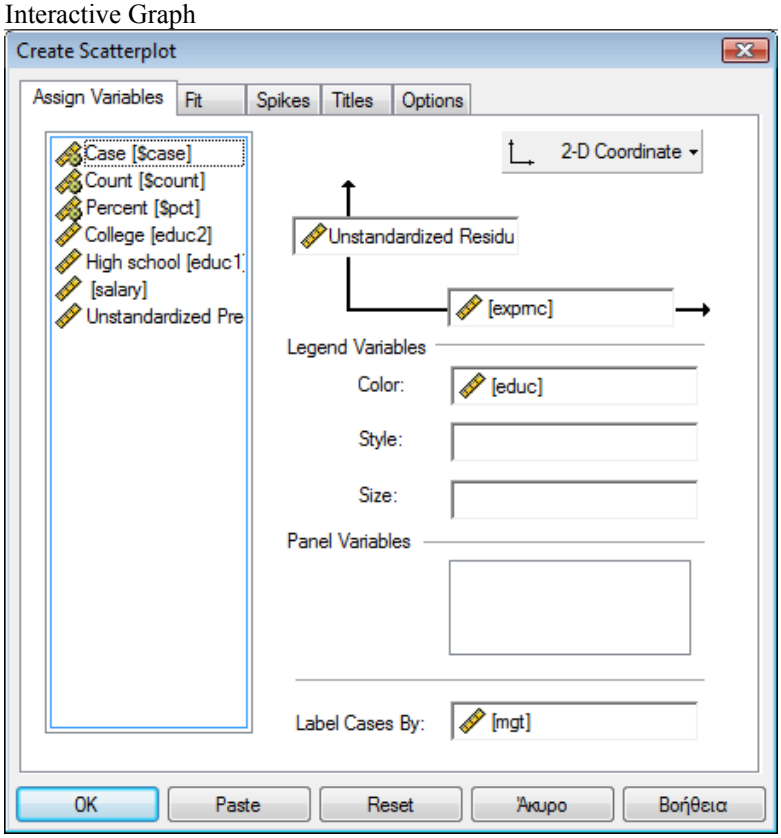

Unstandardized Residual
2.000,00000
1.000,00000
0,00000
-1.000,00000
-2.000,00000
0,00
5,00
10,00
15,00
20,00
exprnc

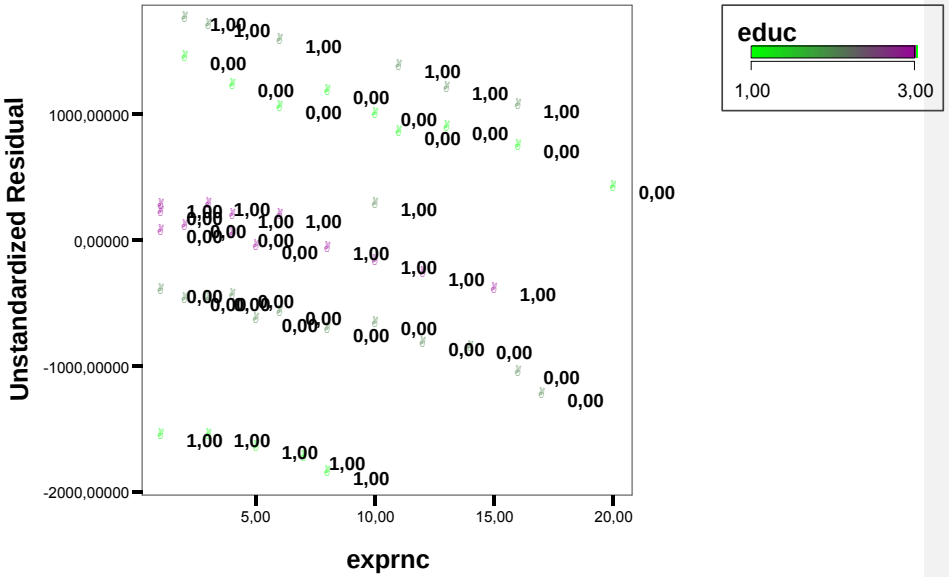
Τι παρατηρεί κάποιος? Τρία επίπεδα υπολοίπων.

The plot suggests that there may be three or more
specific levels of residuals. Possibly the indicator variables that have been defined are not adequate for explaining the effects of education and management status. Actually, each residual is identified with one of the six education-management combinations. To see this we plot the residuals against Category (a new categorical variable that takes a separate value for each of the six combinations). This graph is, in effect, a plot of residuals versus a potential predictor variable that has not yet been used in the equation. The

Αναρωτιόμαστε τότε τι ρόλο μπορεί να παίζουν οι δείκτριες?

Interactive Graph





Είναι φανερό ότι παίζουν ρόλο οι συνδυασμοί των δύο κατηγορικών μεταβλητών. Μπορεί κάποιος να δημιουργήσει μία νέα μεταβλητή που λαμβάνει μία από τις 6 τιμές των συνδυασμών.

For example, the effect of a management position is measured as 61, independently of the level of educational attainment. The nonadditive effects of these variables can be evaluated by constructing additional variables that are used to measure what may be referred to as *multiplicative* or *interaction effects.* Interaction variables are defined as products of the existing indicator variables *(El . M)* and *(E2. M)* . The inclusion of these two variables on the right-hand side of (5.1) leads to a model that is no longer additive in education and management, but recognizes the multiplicative effect of these two variables

Προσαρμόζω το μοντέλο

Coefficients(a)

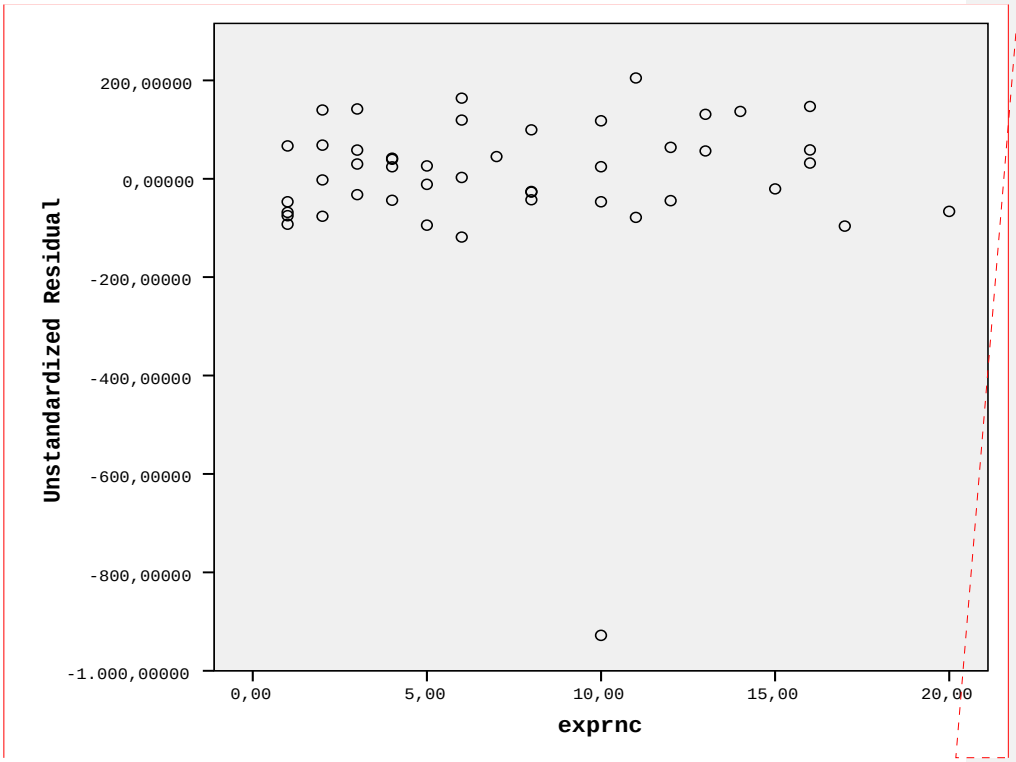
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	11203,434	79,065		141,698	,000
	exprnc	496,987	5,566	,545	89,283	,000
	High school	-1730,748	105,334	-,171	-16,431	,000
	College	-349,078	97,568	-,037	-3,578	,001
	mgt	7047,412	102,589	,749	68,695	,000
	educ1mgt	-3066,035	149,330	-,205	-20,532	,000
	educ2mgt	1836,488	131,167	,141	14,001	,000

a Dependent Variable: salary

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,999(a)	,999	,999	173,80861

a Predictors: (Constant), educ2mgt, exprnc, educ1mgt, College, High school, mgt
b Dependent Variable: salary



Σχόλιο [a6]: πιθανή ακραία τιμή !!!! Show Data Value πρόκειται για την 33 παρατήρηση!!!

Όπως θα δούμε σε επόμενη ενότητα ελέγχουμε τις ακραίες:

- Παρατηρήσεις με απόλυτες τιμές των τυποποιημένων υπολοίπων μεγαλύτερες του 3 θεωρούνται ακραίες.

Έχει -5,33993 οπότε είναι ακραία. Δεν την διώχνουμε αυτόματα….Βλέπουμε τι ισχύει για αυτήν….τι πληροφορίες μας δίνει…Άτομο με 10 ετη εμπειρίας Educ 3 MGT 1 και μισθό 23174 ενώ η πρόβλεψη 24708 (υπερεκτίμηση). Αφαιρούμε αυτή την παρατήρηση και προσαρμόζουμε ξανά το μοντέλο! (Data Select cases (observation 33 excluded)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	11199,714	30,533		366,802	,000
	exprnc	498,418	2,152	,557	231,640	,000
	High school	-1741,336	40,683	-,175	-42,803	,000
	College	-357,042	37,681	-,038	-9,475	,000
	mgt	7040,580	39,619	,754	177,707	,000
	educ1mgt	-3051,763	57,674	-,208	-52,914	,000

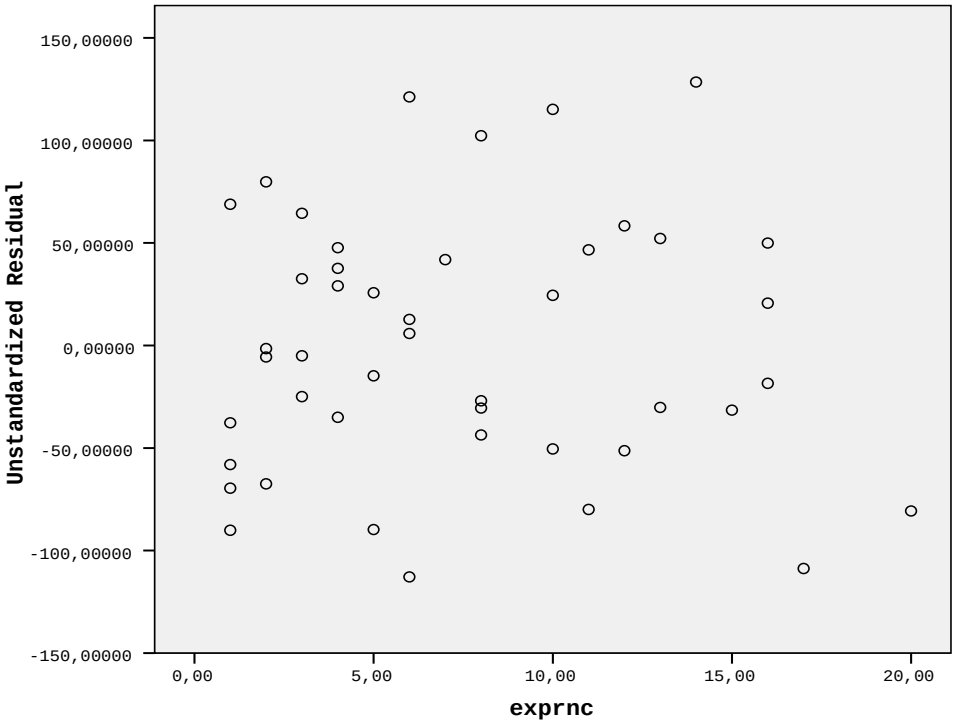
	educ2mgt	1997,531	51,785	,147	38,574	,000

a Dependent Variable: salary

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1,000(a)	1,000	1,000	67,11893

a Predictors: (Constant), educ2mgt, exprnc, educ1mgt, College, High school, mgt
b Dependent Variable: salary



Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
1	(Constant)	11199,714	30,533		366,802	,000	11137,902	11261,525
	exprnc	498,418	2,152	,557	231,640	,000	494,06	
	High school	-1741,336	40,683	-,175	-42,803	,000	-1823,693	-1658,979
	College	-357,042	37,681	-,038	-9,475	,000	-433,324	-280,761
	mgt	7040,580	39,619	,754	177,707	,000	6960,376	7120,785
	educ1mgt	-3051,763	57,674	-,208	-52,914	,000	-3168,519	-2935,008
	educ2mgt	1997,531	51,785	,147	38,574	,000	1892,697	2102,364

a Dependent Variable: salary

Σχόλιο [a7]: Το δ.ε.

Using a regression model with dummy variable and interaction terms it has been possible to explain all the variation in salaries of computer professionals selected for this survey.

Ιδια αποτελέσματα με το να προσαρμόσουμε το μοντέλο με 5 δείκτριες. Ποιο είναι προτιμότερο?
Το πρώτο γιατι δίνει τη δυνατότητα separate the effects of the three sets of explanatory variables α) education b)management c) education-management.

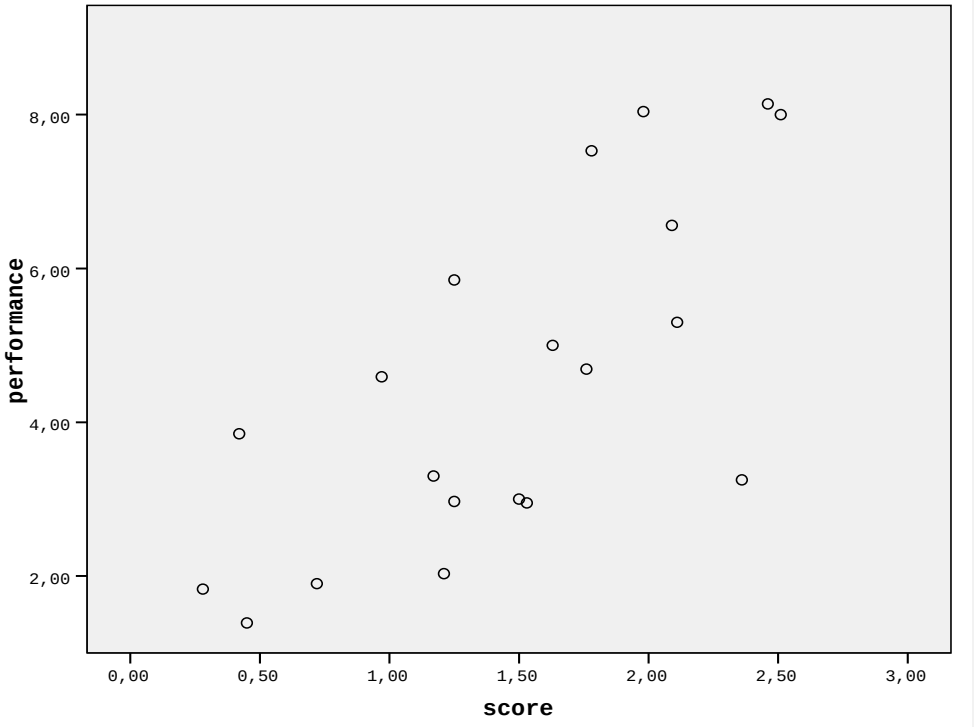
Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	11199,714	30,533		366,802	,000
	exprnc	498,418	2,152	,557	231,640	,000
	educ1mgt1	2247,481	42,713	,153	52,618	,000
	educ1mgt0	-1741,336	40,683	-,151	-42,803	,000
	educ2mgt1	8681,068	42,579	,640	203,881	,000
	educ2mgt0	-357,042	37,681	-,034	-9,475	,000
	educ3mgt1	7040,580	39,619	,583	177,707	,000

a Dependent Variable: salary

Στο αρχείο δεδομένων dummy2.sav καταγράφονται η απόδοση στην εργασία (Job performance), η τιμή σε ένα τεστ δεξιοτήτων πριν την πρόσληψη (score) και το αν ανήκει το άτομο σε μειονοτική ομάδα ή όχι. Να εξετάσετε αν και πως μπορεί να «στηθεί» ένα μοντέλο πρόβλεψης της απόδοσης στην εργασία

<u>Σχόλια για το αρχείο dummy2.sav</u>



Model 1:

Pooled $Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{ij}$

Model 2 separate $Y_{1j} = \beta_{10} + \beta_{11} X_{1j} + u_{ij}$

$Y_{2j} = \beta_{20} + \beta_{21} X_{2j} + u_{ij}$

Τι θα ελέγξουμε?

$\beta_{11} = \beta_{21}, \beta_{10} = \beta_{22}$

ΤΕΣΤ Of Coincidence

Ποια η χρησιμότητα του ελέγχου? Να θέτουμε ίδιους όρους για την πρόσληψη ή όχι?
Τι ξέρουμε από τη θεωρία?
Βλέπε Seber p. 201

Το μοντέλο δύο μπορεί να γραφεί (full):
Υπό την υπόθεση ίσων διακυμάνσεων στα 2 γκρουπ άρα ένα τεστ του Levene

$$\begin{pmatrix} \mathrm{Y}_{11} \\ \mathrm{Y}_{12} \\ \vdots \\ \mathrm{Y}_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \mathrm{X}_{11} & 0 \\ 1 & 0 & \mathrm{X}_{12} & 0 \\ & & & \\ 1 & 0 & \mathrm{X}_{1n_1} & 0 \\ 0 & 1 & 0 & X_{21} \\ 0 & 1 & 0 & X_{22} \\ & & & \\ 0 & 0 & 0 & X_{2n_2} \end{pmatrix} \begin{pmatrix} \beta_{10} \\ \beta_{20} \\ \beta_{11} \\ \beta_{21} \end{pmatrix} + u$$

Τεστ of coincidence σημαίνει

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_{10} \\ \beta_{20} \\ \beta_{11} \\ \beta_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Τότε προκύπτει το μοντέλο (reduced):

$$\begin{pmatrix} \mathrm{Y}_{11} \\ \mathrm{Y}_{12} \\ \vdots \\ \mathrm{Y}_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{X}_{11} \\ 1 & \mathrm{X}_{12} \\ & \\ 1 & \mathrm{X}_{1n_1} \\ 1 & X_{21} \\ 1 & X_{22} \\ & \\ 1 & X_{2n_2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \\ \beta_1 \\ \\ \end{pmatrix} + u$$

Το στατιστικό για τον έλεγχο είναι το

$$F = \frac{\{RSS(reduced) - RSS(\,full)\}/(2k-2)}{RSS(\,full)/(N-2k)} = \frac{\{RSS(reduced) - RSS(\,full)\}/2}{RSS(\,full)/16}$$

Με κρίσιμη περιοχή $F \geq F_{2k-2,n-2k,a} = IDF.F(1-a, 2k-2, n-2k)$ και p-value $1 - CDF.F(F, 2k-2, n-2k)$

Αρκεί να υπολογιστούν αυτά τα δύο
Full model is equivalent with:
Γιατί ?Γιατί έτσι όπως ήταν γραμμένο έχει πρόβλημα πολυσυγγραμμικότητας

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & X_{11} & X_{11} \\ 1 & 1 & X_{12} & X_{12} \\ & & & \\ 1 & 1 & X_{1n_1} & X_{2n_2} \\ 1 & 0 & X_{21} & 0 \\ 1 & 0 & X_{22} & 0 \\ & & & \\ 1 & 0 & X_{2n_2} & 0 \end{pmatrix} \begin{pmatrix} a \\ \beta \\ \gamma \\ \delta \end{pmatrix} + u$$

Όπου απλά $\alpha + \beta = \beta_{10}, \gamma + \delta = \beta_{11}, \gamma = \beta_{21}, \alpha = \beta_{11}$
Αυτό το μοντέλο προσαρμόζεται πολύ εύκολα
Τοτε προκύπτει ότι

Full
Constant, minority, score, score*minority

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	2,010	1,050		1,914	,074
	minority	-1,913	1,540	-,441	-1,242	,232
	score	1,313	,670	,400	1,959	,068
	scoreminority	1,998	,954	,820	2,093	,053

a Dependent Variable: performance

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	62,636	3	20,879	10,553	,000(a)
	Residual	31,655	16	1,978		
	Total	94,291	19			

a Predictors: (Constant), scoreminority, score, minority
b Dependent Variable: performance

Reduced

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48,723	1	48,723	19,246	,000(a)
	Residual	45,568	18	2,532		
	Total	94,291	19			

a Predictors: (Constant), score
b Dependent Variable: performance

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	1,035	,868		1,192	,249
	score	2,361	,538	,719	4,387	,000

a Dependent Variable: performance

Επομένως είναι

$$F = \frac{(45{,}568\text{-}31{,}655)/2}{31{,}655/16} = \frac{6.9565}{1.9784375} = 3.516 \text{ p-value} \qquad ,008<0.05$$

Άρα διαφορετικά μοντέλα….

white

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	,097	1,035		,094	,928
	score	3,311	,624	,882	5,305	,001

a Dependent Variable: performance

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,542(a)	,293	,205	1,51239

a Predictors: (Constant), score

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	2,010	1,129		1,780	,113
	score	1,313	,721	,542	1,822	,106

a Dependent Variable: performance

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,882(a)	,779	,751	1,29213

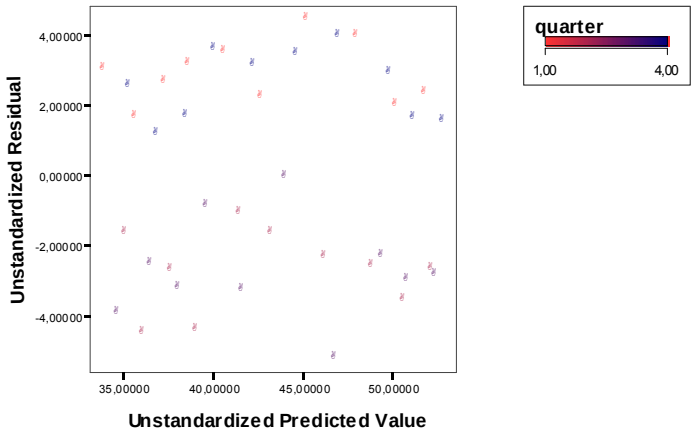
a Predictors: (Constant), score

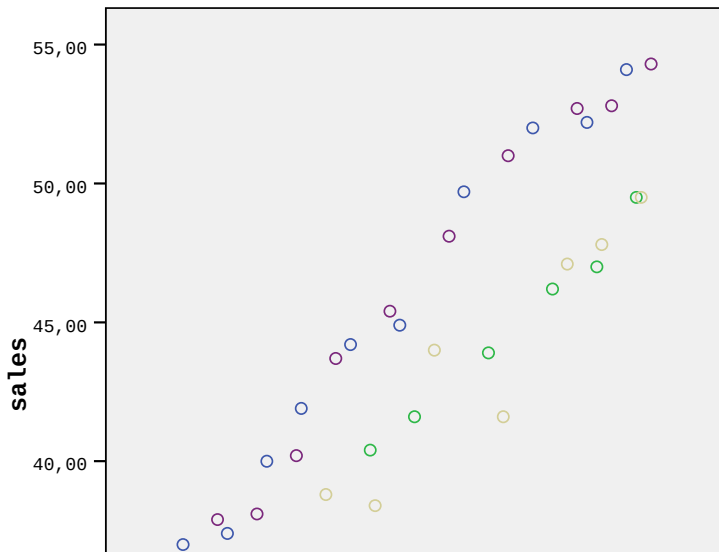
τα οποία θα πρέπει να εξετάσουμε

Σχόλια για το αρχείο dummy3.sav

Χρήση δείκτριας για Piecewise regression.
Εξήγηση θεωρίας
Εισαγωγή της (χ-κ)*δεικτρια

Σχόλια για το αρχείο autocorrelation3.sav
Το αφήνουν ως άσκηση




55,00
50,00
45,00
40,00
sales