

Άσκηση 11. Δίνονται οι παρακάτω παρατηρήσεις:

X_1	X_2	X_3	X_4	Y
44	47	50	62	74
31	43	46	61	96
37	30	51	42	55
39	44	45	59	31
36	23	42	31	128
28	43	35	40	59
37	17	48	18	145
15	14	20	37	69
39	46	46	71	63
49	45	71	62	82
45	16	69	43	183
20	32	33	50	66
21	10	33	20	150
48	11	59	23	176
38	17	60	17	187
26	42	45	47	88
39	14	48	11	150
30	35	42	37	98
37	24	51	42	98
13	37	28	68	65
24	14	36	32	150
10	14	17	21	94
33	43	57	56	72
13	28	20	40	80
12	11	31	26	103
30	32	41	24	94
21	37	43	42	83
26	26	45	22	123
25	23	33	34	98
42	33	47	43	82

1) Να βρεθεί ο πίνακας συσχέτισης μεταξύ των X_1, X_2, X_3, X_4 . Ποιά ζεύγη μεταβλητών παρουσιάζουν υψηλή συσχέτιση;

2) i) Να εξεταστούν όλα τα δυνατά γραμμικά μοντέλα που περιέχουν τις X_1, X_2, X_3, X_4 και την Y . Να βρεθεί το καλύτερο με βάση: α) το R^2 , β) το $R^2(\text{adj})$, γ) το MSE_p και δ) το C_p .

ii) Να βρεθεί το καλύτερο μοντέλο με βάση τις διαδικασίες Stepwise regression και Backward elimination και να περιγραφούν όλα τα ενδιάμεσα βήματα.

3) **Στο πλήρες μοντέλο:** i) Εκτιμήστε τα β_i , σ^2 και υπολογίστε τους συντελεστές προσδιορισμού, R^2 , $R^2(\text{adj})$.

ii) Βρείτε 95% διάστημα εμπιστοσύνης για κάθε ένα από τα β_i . Βρείτε από κοινού δ.ε. 95% για τα $\beta_1, \beta_2, \beta_3, \beta_4$ με τη μέθοδο Bonferroni.

iii) Να κάνετε τους ελέγχους $H_0: \beta_i = 0$ με $H_1: \beta_i \neq 0$, $i=1,2,3,4$ και $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ με $H_1: \beta_1 \neq 0$ ή $\beta_2 \neq 0$ ή $\beta_3 \neq 0$ ή $\beta_4 \neq 0$ σε ε.σ. 1%.

iv) Βρείτε 95% δ.ε. για τη μέση και ατομική πρόβλεψη του Y όταν $X_1 = X_2 = X_3 = X_4 = 30$.

v) Κάνετε τον έλεγχο $H_0: \beta_2 = -\beta_3$ με $H_1: \beta_2 \neq -\beta_3$ σε ε.σ. $\alpha = 5\%$. Βρείτε δ.ε. 95% για το $\beta_2 + \beta_3$.

Λύση. 1) Χρησιμοποιώντας τη διαδικασία /Analyze/correlate/Bivariate προκύπτει ο πίνακας

		Correlations			
		X1	X2	X3	X4
X1	Pearson Correlation	1,000	,224	,886**	,147
	Sig. (2-tailed)		,235	,000	,437
	N	30	30	30	30
X2	Pearson Correlation	,224	1,000	,209	,814**
	Sig. (2-tailed)	,235		,267	,000
	N	30	30	30	30
X3	Pearson Correlation	,886**	,209	1,000	,151
	Sig. (2-tailed)	,000	,267		,426
	N	30	30	30	30
X4	Pearson Correlation	,147	,814**	,151	1,000
	Sig. (2-tailed)	,437	,000	,426	
	N	30	30	30	30

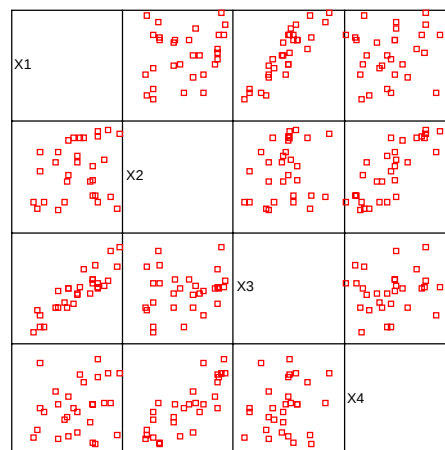
** . Correlation is significant at the 0.01 level (2-tailed).

και άρα ο πίνακας συσχέτισης μεταξύ των X_1, X_2, X_3, X_4 είναι

	X1	X2	X3	X4
X1	1,000			
X2	,224	1,000		
X3	,886	,209	1,000	
X4	,147	,814	,151	1,000

(Υπενθυμίζεται ότι ο πίνακας που προκύπτει από την διαδικασία Linear Regression/Statistics: Regression Coefficients, covariance Matrix είναι ο πίνακας συσχέτισης και συνδιασποράς μεταξύ των $\hat{\beta}_i$ και όχι μεταξύ των στηλών X_i).

Παρατηρούμε ότι τα ζεύγη X1,X3 και X2,X4 παρουσιάζουν υψηλή συσχέτιση. Αυτό γίνεται εύκολα αντιληπτό και από το παρακάτω γράφημα (/Graphs/Scatter/Matrix).



Επομένως (για λόγους που έχουμε αναπτύξει στην άσκ. 9) ίσως θα πρέπει στο μοντέλο να μείνει μόνο μία από τις X1,X3 και μία από τις X2,X4. Σε αυτό το σημείο αξίζει να παρατηρήσουμε ότι δεν έχει νόημα να επικαλεστούμε τα p-value που δίνονται στον παραπάνω πίνακα διότι αυτά προκύπτουν υποθέτοντας ότι οι X_i είναι τυχαίες παρατηρήσεις από κάποια κατανομή κάτι που δεν συμβαίνει στην προκειμένη περίπτωση (στο γραμμικό μοντέλο θεωρούνται ως ερμηνευτικές μεταβλητές, δηλαδή μη-τυχαίες).

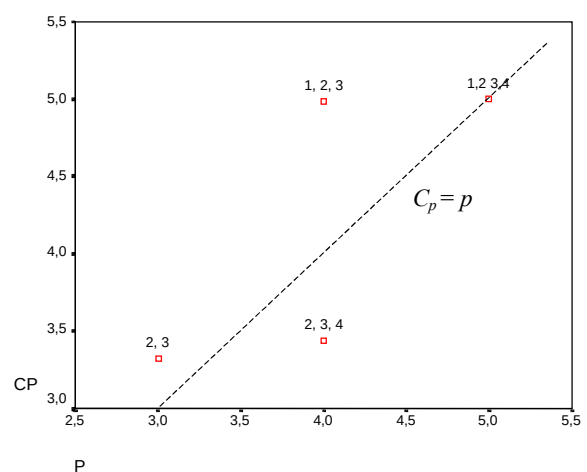
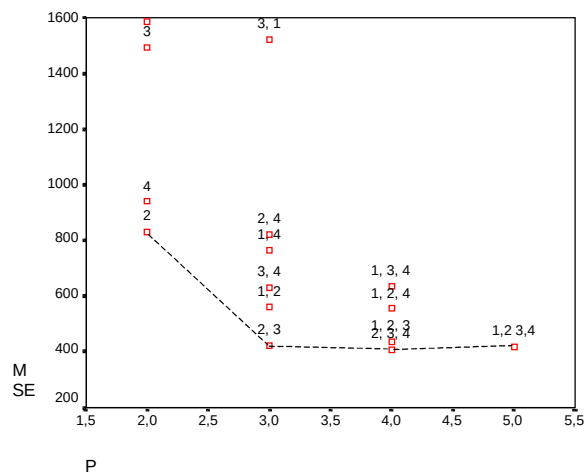
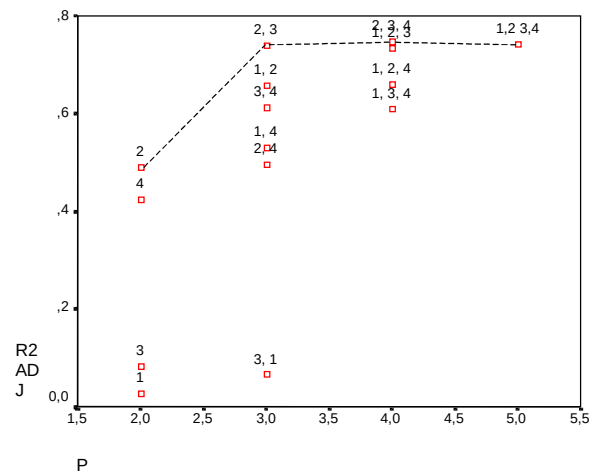
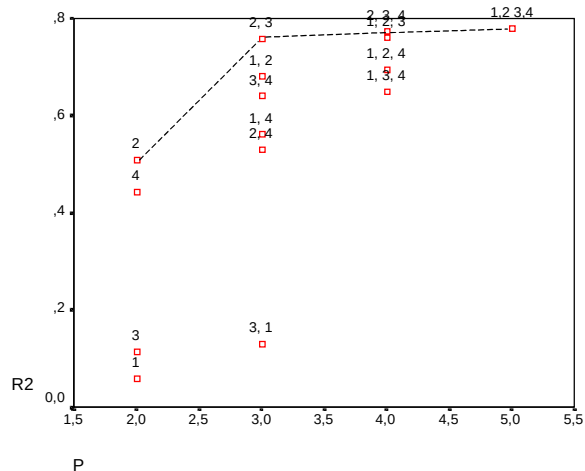
2) Εξετάζοντας όλα τα δυνατά μοντέλα λαμβάνουμε τον πίνακα.

Model	p	R ²	R ² (Adj)	MSE _p	C _p
X1	2	0.059	0.026	1586.87	80.08
X2	2	0.509	0.491	828.98	29.42
X3	2	0.115	0.083	1493.72	73.86
X4	2	0.443	0.423	939.99	36.84
X1, X2	3	0.680	0.656	559.88	12.09
X1, X3	3	0.129	0.065	1523.18	74.19
X1, X4	3	0.562	0.530	765.78	25.36
X2, X3	3	0.758	0.740	423.89	3.32
X2, X4	3	0.530	0.495	822.22	29.00
X3, X4	3	0.640	0.613	629.81	16.60
X1, X2, X4	4	0.695	0.659	554.77	12.44
X1, X2, X3	4	0.761	0.733	434.86	4.99
X1, X3, X4	4	0.650	0.610	635.86	17.47
X2, X3, X4	4	0.774	0.748	409.90	3.44
X1, X2, X3, X4	5	0.778	0.743	418.85	5.00

Οπότε,

- 1) Σύμφωνα με το R^2 καλύτερο θεωρείται το μοντέλο X2, X3 («σταθεροποιεί» το R^2)
- 2) Σύμφωνα με το $R^2(\text{adj})$ καλύτερο θεωρείται το μοντέλο X2, X3, X4 (μεγαλύτερο $R^2(\text{adj})$)
- 3) Σύμφωνα με το MSE_p καλύτερο θεωρείται το μοντέλο X2, X3, X4 (μικρότερο MSE_p)
- 4) Σύμφωνα με το C_p καλύτερο θεωρείται το μοντέλο X2, X3 (μικρότερο C_p με $C_p \approx p$).

Υπενθυμίζεται ότι τα κριτήρια με βάση τα $R^2(\text{adj})$ και MSE_p είναι ισοδύναμα διότι $R^2(\text{adj}) = 1 - (n-1) \text{MSE}_p / \text{SST}$. Τα παραπάνω γίνονται φανερά και από τα παρακάτω γραφήματα:



Σύμφωνα με παραπάνω παρατήρηση, επειδή η X4 παρουσιάζει υψηλή συσχέτιση με την X2 τελικά είναι προτιμότερο το μοντέλο X2, X3 και όχι το X2, X3, X4 (η εξάρτηση της Y από την X4 ίσως να είναι φαινομενική και να οφείλεται στο ότι η Y εξαρτάται από την X2 ή οποία παρουσιάζει υψηλή συσχέτιση με την X4).

ii) Stepwise regression

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	X2	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	X3	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,713 ^a	,509	,491	28,79
2	,870 ^b	,758	,740	20,59

a. Predictors: (Constant), X2

b. Predictors: (Constant), X2, X3

ANOVA^c

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24029,716	1	24029,716	28,987	,000 ^a
	Residual	23211,484	28	828,982		
	Total	47241,200	29			
2	Regression	35796,122	2	17898,061	42,223	,000 ^b
	Residual	11445,078	27	423,892		
	Total	47241,200	29			

a. Predictors: (Constant), X2

b. Predictors: (Constant), X2, X3

c. Dependent Variable: Y

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	166,871	13,248		12,596	,000
	X2	-2,308	,429	-,713	-5,384	,000
2	(Constant)	109,091	14,492		7,528	,000
	X2	-2,654	,313	-,820	-8,466	,000
	X3	1,569	,298	,510	5,269	,000

a. Dependent Variable: Y

Το καλύτερο μοντέλο με βάση τη διαδικασία Stepwise regression είναι το X2, X3.

Backward elimination

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	X4, X1, X2, X3	,	Enter
2	,	X1	Backward (criterion: Probability of F-to-remove $\geq$,100).
3	,	X4	Backward (criterion: Probability of F-to-remove $\geq$,100).

a All requested variables entered.

b Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,882 ^a	,778	,743	20,47
2	,880 ^b	,774	,748	20,25
3	,870 ^c	,758	,740	20,59

a. Predictors: (Constant), X4, X1, X2, X3

b. Predictors: (Constant), X4, X2, X3

c. Predictors: (Constant), X2, X3

ANOVA^d

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36769,989	4	9192,497	21,947	,000 ^a
	Residual	10471,211	25	418,848		
	Total	47241,200	29			
2	Regression	36583,703	3	12194,568	29,750	,000 ^b
	Residual	10657,497	26	409,904		
	Total	47241,200	29			
3	Regression	35796,122	2	17898,061	42,223	,000 ^c
	Residual	11445,078	27	423,892		
	Total	47241,200	29			

a. Predictors: (Constant), X4, X1, X2, X3

b. Predictors: (Constant), X4, X2, X3

c. Predictors: (Constant), X2, X3

d. Dependent Variable: Y

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	113,215	15,065		7,515	,000
	X1	-,492	,737	-,136	-,667	,511
	X2	-2,029	,533	-,627	-3,804	,001
	X3	1,923	,626	,625	3,072	,005
	X4	-,564	,399	-,229	-1,412	,170
2	(Constant)	114,515	14,778		7,749	,000
	X2	-2,065	,525	-,638	-3,935	,001
	X3	1,555	,293	,506	5,307	,000
	X4	-,547	,394	-,222	-1,386	,177
3	(Constant)	109,091	14,492		7,528	,000
	X2	-2,654	,313	-,820	-8,466	,000
	X3	1,569	,298	,510	5,269	,000

a. Dependent Variable: Y

Το καλύτερο μοντέλο με βάση τη διαδικασία Backward elimination είναι και πάλι το X2, X3.

3) Θεωρούμε το πλήρες μοντέλο:

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	X4, X1, X2, X3 ^a		Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,882 ^a	,778	,743	20,47

a. Predictors: (Constant), X4, X1, X2, X3

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36769,989	4	9192,497	21,947	,000 ^a
	Residual	10471,211	25	418,848		
	Total	47241,200	29			

a. Predictors: (Constant), X4, X1, X2, X3

b. Dependent Variable: Y

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	113,215	15,065		7,515	,000	82,187	144,242
X1	-,492	,737	-,136	-,667	,511	-2,010	1,027
X2	-2,029	,533	-,627	-3,804	,001	-3,127	-,930
X3	1,923	,626	,625	3,072	,005	,634	3,213
X4	-,564	,399	-,229	-1,412	,170	-1,387	,259

a. Dependent Variable: Y

ii) Τα δ.ε. 95% για κάθε ένα από τα β_i δίνονται στον παραπάνω πίνακα. Συγκεκριμένα,

95% Confidence Interval for B		
	Lower Bound	Upper Bound
β_0	82,187	144,242
β_1	-2,010	1,027
β_2	-3,127	-,930
β_3	,634	3,213
β_4	-1,387	,259

Για την εύρεση από κοινού δ.ε. $1-\alpha = 95\%$ για τα $\beta_1, \beta_2, \beta_3, \beta_4$ με τη μέθοδο Bonferroni αρκεί να πάρουμε ως $I_i = (L_i, U_i)$ ένα δ.ε. για το β_i συντελεστού $1-\alpha/4, i = 1, 2, 3, 4$ διότι τότε

$$\begin{aligned}
 P((\beta_1, \beta_2, \beta_3, \beta_4) \in I_1 \times I_2 \times I_3 \times I_4) &= P(\beta_i \in I_i, i = 1, 2, 3, 4) = P(A_1 \cap \dots \cap A_4) \\
 &= 1 - P(A_1^C \cup \dots \cup A_4^C) \geq 1 - P(A_1^C) - \dots - P(A_4^C) \\
 &= 1 - P(\beta_1 \notin I_1) - \dots - P(\beta_4 \notin I_4) = 1 - 4 \cdot \frac{\alpha}{4} = 1 - \alpha
 \end{aligned}$$

Επομένως, δ.ε. για τα β_i συντελεστού $1-\alpha/4 = 98.75\%$ είναι:

$$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_{n-p}((\alpha/4)/2) = -0.492 \pm 0.737 \cdot t_{25}(0.00625) = (-2.467, 1.483)$$

$$\hat{\beta}_2 \pm s(\hat{\beta}_2) t_{n-p}((\alpha/4)/2) = -2.029 \pm 0.533 \cdot t_{25}(0.00625) = (-3.457, -0.600)$$

$$\hat{\beta}_3 \pm s(\hat{\beta}_3) t_{n-p}((\alpha/4)/2) = 1.923 \pm 0.626 \cdot t_{25}(0.00625) = (0.245, 3.600)$$

$$\hat{\beta}_4 \pm s(\hat{\beta}_4) t_{n-p}((\alpha/4)/2) = -0.564 \pm 0.399 \cdot t_{25}(0.00625) = (-1.633, 0.505)$$

($t_{25}(0.00625) = \text{IDF.T}(0.99375, 27) = 2,68$). Άρα η περιοχή εμπιστοσύνης 95% για το $(\beta_1, \beta_2, \beta_3, \beta_4)$ θα είναι η

$$(-2.467, 1.483) \times (-3.457, -0.600) \times (0.245, 3.600) \times (-1.633, 0.505).$$

iii) Τα p-values των 4 πρώτων ελέγχων $H_0: \beta_i = 0$ με $H_1: \beta_i \neq 0, i = 1, 2, 3, 4$ δίνονται στον πίνακα των coefficients (0.511, 0.001, 0.005, 0.170). Επομένως, απορρίπτουμε τις υποθέσεις $H_0: \beta_2 = 0, H_0: \beta_3 = 0$.

Για τον από κοινού έλεγχο ως γνωστό χρησιμοποιούμε το F-test από τον πίνακα ANOVA. Το αντίστοιχο p-value είναι 0.000 οπότε, όπως ήταν αναμενόμενο, απορρίπτουμε την $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

iv) Εισάγοντας μία επιπλέον γραμμή στα δεδομένα και εκτελώντας τη διαδικασία της γραμμικής παλινδρόμησης (save Prediction intervals) προκύπτει ότι

δ.ε. 95% για μέση πρόβλεψη του Y ($X_1 = X_2 = X_3 = X_4 = 30$) : (58.25021, 98.49249),

δ.ε. 95% για ατομική πρόβλεψη του Y ($X_1 = X_2 = X_3 = X_4 = 30$) : (31.66493, 125.0778)

v) Θα χρησιμοποιήσουμε τον μετασχηματισμό: $\gamma = \beta_2 + \beta_3 \Leftrightarrow \beta_3 = \gamma - \beta_2$ και το μοντέλο γίνεται

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + (\gamma - \beta_2) X_3 + \beta_4 X_4 + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 (X_2 - X_3) + \gamma X_3 + \beta_4 X_4 + \varepsilon$$

από όπου προκύπτει ο πίνακας ($X_{23} = X_2 - X_3$)

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	113,215	15,065		7,515	,000	82,187	144,242
X1	-,492	,737	-,136	-,667	,511	-2,010	1,027
X23	-2,029	,533	-,809	-3,804	,001	-3,127	-,930
X3	-,106	,830	-,034	-,127	,900	-1,815	1,604
X4	-,564	,399	-,229	-1,412	,170	-1,387	,259

a. Dependent Variable: Y

Επειδή p-value = 0.9 δεν απορρίπτουμε ότι $H_0: \gamma = 0 \Leftrightarrow \beta_2 = -\beta_3$. Το δ.ε. 95% για το $\gamma = \beta_2 + \beta_3$ είναι (-1.815, 1.604).