

Άσκηση 8. Δίνονται οι παρακάτω 30 παρατηρήσεις (πίνακας Α) με βάση τις οποίες θέλουμε να δημιουργήσουμε ένα γραμμικό μοντέλο για την πρόβλεψη της Y μέσω των ανεξάρτητων μεταβλητών X_1, X_2 .

Πίνακας Α

i	Y	X_1	X_2
1	233	68	16
2	443	66	77
3	199	63	28
4	444	7	89
5	303	90	23
6	256	39	30
7	507	58	95
8	226	12	39
9	216	71	9
10	305	92	29
11	490	87	67
12	481	87	78
13	353	25	83
14	216	56	40
15	523	49	95
16	328	67	39
17	365	91	40
18	227	51	28
19	369	2	84
20	370	64	67
21	162	15	34
22	538	93	82
23	543	99	90
24	91	8	13
25	461	78	83
26	207	86	15
27	115	50	13
28	68	29	25
29	170	73	6
30	438	87	59

Πίνακας Β

i	Y	X_1	X_2
1	11.9	19.5	43.1
2	22.8	24.7	49.8
3	18.7	30.7	51.9
4	20.1	29.8	54.3
5	12.9	19.1	42.2
6	21.7	25.6	53.9
7	27.1	31.4	58.5
8	25.4	27.9	52.1
9	21.3	22.1	49.9
10	19.3	25.5	53.5
11	25.4	31.1	56.6
12	27.2	30.4	56.7
13	11.7	18.7	46.5
14	17.8	19.7	44.2
15	12.8	14.6	42.7
16	23.9	29.5	54.4
17	22.6	27.7	55.3
18	25.4	30.2	58.6
19	14.8	22.7	48.2
20	21.1	25.2	51.0

1. Θεωρώντας ότι μοντέλο είναι

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \mathbf{I}_n),$$

- i) Εκτιμήστε τα $\beta_0, \beta_1, \beta_2, \sigma^2$ και υπολογίστε τους συντελεστές προσδιορισμού, $R^2, R^2(\text{adj})$.
- ii) Βρείτε τον πίνακα συσχέτισης και τον πίνακα διασποράς των $\hat{\beta}_1, \hat{\beta}_2$.
- iii) Βρείτε 95% διάστημα εμπιστοσύνης για κάθε ένα από τα $\beta_0, \beta_1, \beta_2$. Βρείτε από κοινού δ.ε. 95% για τα β_1, β_2 , με τη μέθοδο Bonferroni.
- iv) Να κάνετε τους ελέγχους $H_0: \beta_1=0$ με $H_1: \beta_1 \neq 0$, $H_0: \beta_2=0$ με $H_1: \beta_2 \neq 0$ και $H_0: \beta_1=\beta_2=0$ με $H_1: \beta_1 \neq 0$ ή $\beta_2 \neq 0$ σε ε.σ. 1%.
- v) Βρείτε 95% δ.ε. για τη μέση και ατομική πρόβλεψη του Y όταν $X_1=50, X_2=60$.
- vi) Κάνετε τον έλεγχο $H_0: 2\beta_1=\beta_2$ με $H_1: 2\beta_1 \neq \beta_2$ σε ε.σ. $\alpha = 5\%$. Βρείτε δ.ε. 95% για το $2\beta_1-\beta_2$.
- vii) Ποιές παρατηρήσεις επηρεάζουν περισσότερο το μοντέλο ($p_{ii} > 3 \cdot 2/n$) και ποιες θεωρούνται ασυνήθιστες ($|\hat{\varepsilon}_i^*| > 2$); Να γίνει έλεγχος κανονικότητας των παρατηρήσεων ($K-S$ για τα κατάλοιπα).
- viii) Το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ βελτιώνει ουσιαστικά το προηγούμενο ($\beta_3 = 0$;)

2. Εξετάστε ποιο από τα παρακάτω τρία μοντέλα:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad Y = \beta_0 + \beta_2 X_2 + \varepsilon, \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

είναι το καλύτερο, με βάση το R^2 , το $R^2(\text{adj})$, και το $C_p = \frac{SSE_p}{MSE} - (n - 2p)$.

Παρατηρούμε ότι τα εκτιμημένα β_i δεν αλλάζουν σημαντικά στα τρία μοντέλα. Αυτό συμβαίνει πάντοτε; Εξετάστε γιατί δεν ισχύει κάτι ανάλογο εφαρμόζοντας τα δεδομένα του πίνακα Β.

Άσκηση 9. Δίνονται οι παρακάτω 54 παρατηρήσεις με βάση τις οποίες θέλουμε να δημιουργήσουμε ένα γραμμικό μοντέλο για την πρόβλεψη της Y μέσω κάποιας ή κάποιων από τις ανεξάρτητες μεταβλητές X_1, X_2, X_3, X_4 .

i	X_1	X_2	X_3	X_4	Y
1	6.7	62	81	2.59	200
2	5.1	59	66	1.70	101
3	7.4	57	83	2.16	204
4	6.5	73	41	2.01	101
5	7.8	65	115	4.30	509
6	5.8	38	72	1.42	80
7	5.7	46	63	1.91	80
8	3.7	68	81	2.57	127
9	6.0	67	93	2.50	202
10	3.7	76	94	2.40	203
11	6.3	84	83	4.13	329
12	6.7	51	43	1.86	65
13	5.8	96	114	3.95	830
14	5.8	83	88	3.95	330
15	7.7	62	67	3.40	168
16	7.4	74	68	2.40	217
17	6.0	85	28	2.98	87
18	3.7	51	41	1.55	34
19	7.3	68	74	3.56	215
20	5.6	57	87	3.02	172
21	5.2	52	76	2.85	109
22	3.4	83	53	1.12	136
23	6.7	26	68	2.10	70
24	5.8	67	86	3.40	220
25	6.3	59	100	2.95	276
26	5.8	61	73	3.50	144
27	5.2	52	86	2.45	181
28	11.2	76	90	5.59	574
29	5.2	54	56	2.71	72
30	5.8	76	59	2.58	178
31	3.2	64	65	0.74	71
32	8.7	45	23	2.52	58
33	5.0	59	73	3.50	116
34	5.8	72	93	3.30	295
35	5.4	58	70	2.64	115
36	5.3	51	99	2.60	184
37	2.6	74	86	2.05	118
38	4.3	8	119	2.85	120
39	4.8	61	76	2.45	151
40	5.4	52	88	1.81	148
41	5.2	49	72	1.84	95
42	3.6	28	99	1.30	75
43	8.8	86	88	6.40	483
44	6.5	56	77	2.85	153

45	3.4	77	93	1.48	191
46	6.5	40	84	3.00	123
47	4.5	73	106	3.05	311
48	4.8	86	101	4.10	398
49	5.1	67	77	2.86	158
50	3.9	82	103	4.55	310
51	6.6	77	46	1.95	124
52	6.4	85	40	1.21	125
53	6.4	59	85	2.33	198
54	8.8	78	72	3.20	313

i) Να βρεθεί ο πίνακας συσχέτισης μεταξύ των $X_1, X_2, X_3, X_4, Y' = \log_{10} Y$.

ii) Να εξεταστούν όλα τα δυνατά γραμμικά μοντέλα που περιέχουν τις X_1, X_2, X_3, X_4 και την $Y' = \log_{10} Y$.

Να βρεθεί το καλύτερο με βάση: α) το R^2 , β) το $R^2(\text{adj})$, γ) το MSE_p και δ) το $C_p = \frac{SSE_p}{MSE} - (n - 2p)$.

iii) Να βρεθεί το καλύτερο μοντέλο με βάση τη διαδικασία Stepwise regression και να περιγραφούν όλα τα ενδιάμεσα βήματα.

iv) Να γίνει το ίδιο χρησιμοποιώντας τις διαδικασίες Forward selection και Backward elimination.

Άσκηση 10. Δίνονται οι ετήσιοι μισθοί Y , σε χιλιάδες ευρώ 15 υπαλλήλων που επιλέχτηκαν τυχαία. Επίσης δίνονται τα χρόνια υπηρεσίας X_1 , η ηλικία X_2 και το φύλο τους Φ .

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	15.2	17.6	13.5	12.8	18.0	23.5	19.6	17.0	12.5	12.3	13.5	17.6	15.4	10.5	16.8
X_1	5.4	8.4	6.7	11.2	14.6	22.4	18.3	15.2	17.9	3.5	2.8	9.6	6.8	7.5	13.2
X_2	26	35	42	35	54	28	36	43	48	27	31	44	34	19	26
Φ	Γ	Α	Γ	Γ	Α	Α	Α	Α	Α	Γ	Γ	Γ	Α	Γ	Α

Το μοντέλο για τους άνδρες και τις γυναίκες είναι

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \text{και} \quad Y = c_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

αντίστοιχα. Αν υποθέσουμε ότι οι διασπορές των σφαλμάτων είναι ίσες, εξετάστε αν $\beta_0 = c_0$ σε ε.σ. $\alpha = 0.05$.

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 $\beta_2 + \beta_3$.

Απαντήσεις.

Άσκηση 8. Θεωρούμε το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 \mathbf{I}_n),$$

i) Εκτελούμε τη διαδικασία /Analyze/Regression/Linear με Dependent: Y , Independent: X_1, X_2 (Method: Enter για να εισαχθούν και οι δύο μεταβλητές X_1, X_2 στο μοντέλο). Επίσης επιλέγουμε στα /Statistics και τα εξής: Confidence intervals (δ.ε. για τις εκτιμήσεις των β_i), Covariance Matrix (πίνακας διασποράς των $\hat{\beta}_1, \hat{\beta}_2$), Descriptives (για να εμφανιστούν οι (δειγματικές) μέσες τιμές, τυπικές αποκλίσεις, και συντελεστές συσχέτισης των Y, X_1, X_2). Τα αποτελέσματα από το SPSS είναι:

Descriptive Statistics

	Mean	Std. Deviation	N
Y	321,5667	141,9941	30
X1	58,7667	29,5106	30
X2	49,2000	30,1026	30

Correlations

		Y	X1	X2
Pearson Correlation	Y	1,000	,393	,882
	X1	,393	1,000	-,003
	X2	,882	-,003	1,000
Sig. (1-tailed)	Y	,	,016	,000
	X1	,016	,	,494
	X2	,000	,494	,
N	Y	30	30	30
	X1	30	30	30
	X2	30	30	30

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	X2, X1 ^b	,	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	,966 ^a	,934	,929	37,7795

a. Predictors: (Constant), X2, X1

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	546170,5	2	273085,242	191,331	,000 ^a
	Residual	38536,883	27	1427,292		
	Total	584707,4	29			

a. Predictors: (Constant), X2, X1

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)	4,828	19,369		,249	,805	-34,914	44,570
	X1	1,902	,238	,395	8,002	,000	1,415	2,390
	X2	4,165	,233	,883	17,873	,000	3,687	4,644

a. Dependent Variable: Y

Coefficient Correlations^a

Model		X2	X1
1	Correlations	X2	1,000
		X1	,003
	Covariances	X2	5,431E-02
		X1	1,626E-04

a. Dependent Variable: Y

i) Συνεπώς, οι εκτιμήσεις των $\beta_0, \beta_1, \beta_2, \sigma^2$ είναι

$$\hat{\beta}_0 = 4.828, \hat{\beta}_1 = 1.902, \hat{\beta}_2 = 4.165, \hat{\sigma}^2 = \frac{SSE}{n-p} = MSE = s^2 = 1427.292, \hat{\sigma} = 37.7795$$

Οι συντελεστές προσδιορισμού είναι $R^2 = 0.934, R^2(\text{adj}) = 0.929$.

ii) Ο πίνακας συσχέτισης και ο πίνακας διασποράς των $\hat{\beta}_1, \hat{\beta}_2$ αντίστοιχα είναι:

Coefficient Correlations			
Correlations		X2	X1
	X2	1.000	0.003
	X1	0.003	1.000
Covariances		X2	X1
	X2	54.31	0.16
	X1	0.16	56.51

iii) Τα διαστήματα εμπιστοσύνης 95% για τα $\beta_0, \beta_1, \beta_2$ είναι:

$$\beta_0 : (-34.914, 44.570), \beta_1 : (1.415, 2.390), \beta_2 : (3.687, 4.644).$$

Για την εύρεση από κοινού δ.ε. 95% για τα β_1, β_2 με τη μέθοδο Bonferroni εργαζόμαστε ως εξής: Αναζητούμε μία περιοχή (τετράγωνο) εμπιστοσύνης $(L_1, U_1) \times (L_2, U_2)$ μέσα στο οποίο βρίσκεται το (β_1, β_2) με πιθανότητα τουλάχιστον $1-\alpha = 95\%$. Πιο συγκεκριμένα, αναζητούμε δύο διαστήματα $I_1 = (L_1, U_1), I_2 = (L_2, U_2)$ για τα οποία να ισχύει ότι

$$P((\beta_1, \beta_2) \in I_1 \times I_2) \geq 1-\alpha$$

Αρκεί να πάρουμε ως I_1 ένα δ.ε. για το β_1 συντελεστού $1-\alpha/2$ και ως I_2 ένα δ.ε. για το β_2 συντελεστού $1-\alpha/2$. Αρκεί διότι

$$\begin{aligned} P((\beta_1, \beta_2) \in I_1 \times I_2) &= P(\beta_1 \in I_1, \beta_2 \in I_2) = P(A_1 \cap A_2) = 1 - P(A_1^c \cup A_2^c) \geq \\ &\geq 1 - P(A_1^c) - P(A_2^c) = 1 - P(\beta_1 \notin I_1) - P(\beta_2 \notin I_2) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha \end{aligned}$$

Ένα δ.ε. για το β_1 συντελεστού $1-\alpha/2 = 97.5\%$ είναι:

$$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_{n-p}(0.025/2) = 1.902 \pm 0.238 \cdot t_{27}(0.0125) = (1.334, 2.466)$$

$$\hat{\beta}_2 \pm s(\hat{\beta}_2) t_{n-p}(0.025/2) = 4.165 \pm 0.233 \cdot t_{27}(0.0125) = (3.613, 4.717)$$

($t_{27}(0.0125) = \text{IDF.T}(0.9875, 27) = 2.37$). Άρα η περιοχή εμπιστοσύνης 95% για το (β_1, β_2) θα είναι η $(1.334, 2.466) \times (3.613, 4.717)$.

iv) Για τον έλεγχο $H_0: \beta_1 = 0$ με $H_1: \beta_1 \neq 0$, το πακέτο δίνει p-value 0.000 ($T^* = 8,002$) άρα σε ε.σ. 1% απορρίπτουμε ότι $\beta_1 = 0$. Επίσης, για τον έλεγχο $H_0: \beta_2 = 0$ με $H_1: \beta_2 \neq 0$, το πακέτο δίνει και πάλι p-value 0.000 ($T^* = 17,873$) άρα σε ε.σ. 1% απορρίπτουμε και ότι $\beta_2 = 0$. Τέλος, για τον έλεγχο της υπόθεσης $H_0: \beta_1 = \beta_2 = 0$ με $H_1: \beta_1 \neq 0$ ή $\beta_2 \neq 0$ θα χρησιμοποιήσουμε το F-test του πίνακα ANOVA. Επειδή p-value = 0.000 ($F^* = 191,331$) απορρίπτουμε την H_0 σε ε.σ. 1%.

ν) Για να βρούμε 95% δ.ε. για τη μέση και ατομική πρόβλεψη του Y στο SPSS όταν $X_1=50$, $X_2=60$, προσθέτουμε στην 31^η γραμμή των δεδομένων και στις στήλες των X_1 , X_2 τους αριθμούς 50 και 60 (η 31^η γραμμή στην στήλη του Y αφήνεται κενή). Στη συνέχεια εκτελούμε τη διαδικασία της παλινδρόμησης /Analyze/Regression/Linear επιλέγοντας στο save τα Unstandardized predicted values, Prediction Intervals. Στην 31^η γραμμή λαμβάνονται τα αποτελέσματα:

Αναμενόμενο Y : 349,8760

Δ.ε. 95% για την μέση πρόβλεψη: (334,2195, 365,5324)

Δ.ε. 95% για την μέση πρόβλεψη: (270,7935, 428,9584)

vi) Για να κάνουμε τον έλεγχο $H_0: 2\beta_1=\beta_2$ με $H_0: 2\beta_1\neq\beta_2$ και να βρούμε δ.ε. 95% για το $2\beta_1-\beta_2$ αρκεί να δημιουργήσουμε ένα νέο μοντέλο το οποίο θα περιέχει ως συντελεστή το $\gamma = 2\beta_1-\beta_2$. Συγκεκριμένα, παίρνουμε τον μετασχηματισμό:

$$\beta_0 \rightarrow \beta_0, \beta_1 \rightarrow \beta_1, \beta_2 \rightarrow 2\beta_1 - \gamma$$

από όπου προκύπτει το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + (2\beta_1 - \gamma) X_2 + \varepsilon = \beta_0 + \beta_1 (X_1 + 2X_2) + \gamma(-X_2) + \varepsilon$$

Δημιουργούμε (Transform/compute) δύο νέες μεταβλητές $X_3 := X_1 + 2X_2$, $X_4 := -X_2$. Στη συνέχεια εκτελούμε τη διαδικασία /Analyze/Regression/Linear με Dependent: Y , Independent: X_3 , X_4 λαμβάνοντας τον πίνακα

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	4,828	19,369		,249	,805	-34,914	44,570
X3	1,902	,238	,897	8,002	,000	1,415	2,390
X4	-,361	,529	-,076	-,682	,501	-1,446	,725

a. Dependent Variable: Y

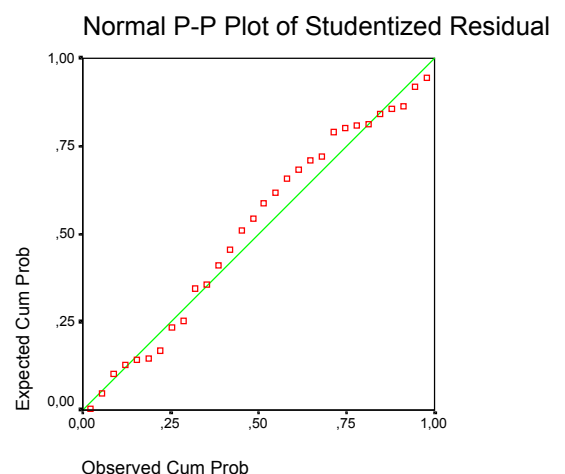
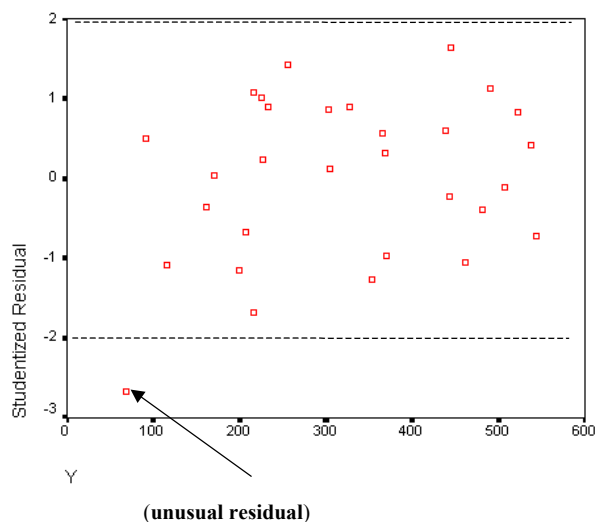
από όπου δεν μπορούμε να απορρίψουμε σε ε.σ. 5% ή 10% ότι $H_0: \gamma = 0$ ή ισοδύναμα ότι $H_0: 2\beta_1 = \beta_2$ διότι το αντίστοιχο p-value του ελέγχου είναι 0.501 ($T^* = -0.682$). Το δ.ε. για το $\gamma = 2\beta_1 - \beta_2$ είναι ίσο με (-1.446, 0.725).

vii) Εκτελώντας τη διαδικασία /Analyze/Regression/Linear με Dependent: Y , Independent: X_1 , X_2 επιλέγουμε στο save τα leverage values και τα studentized residuals. Στον πίνακα των δεδομένων λαμβάνουμε τις στήλες:

i	<i>studentized residuals</i>	<i>leverage</i>
1	0,88685	0,04525
2	-0,22246	0,03153
3	-1,14976	0,01779
4	1,63070	0,16592
5	0,86811	0,06456
6	1,42226	0,02959
7	-0,10924	0,07984
8	1,01500	0,09067
9	1,07771	0,06731
10	0,12093	0,05911
11	1,11799	0,04373
12	-0,39666	0,06331
13	-1,27440	0,08836
14	-1,67170	0,00353
15	0,82343	0,08350
16	0,89845	0,00662

17	0,56320	0,04429
18	0,23163	0,01953
19	0,31107	0,17323
20	-0,96685	0,01316
21	-0,36614	0,08479
22	0,41445	0,08760
23	-0,72417	0,12781
24	0,49283	0,15233
25	-1,05391	0,05827
26	-0,66991	0,07366
27	-1,08271	0,05298
28	-2,66875	0,05753
29	0,03663	0,07890
30	0,60067	0,03528

Παρατηρούμε ότι όλα τα leverages είναι μικρότερα του $3 \cdot 2/n = 0.2$ οπότε δεν υπάρχει κάποια παρατήρηση που να έχει μεγάλη επιρροή στο μοντέλο (αν υπάρχει κάποια τέτοια παρατήρηση, εκτελούμε παλινδρόμηση χωρίς αυτήν για να δούμε αν αλλάζουν σημαντικά οι εκτιμώμενες τιμές των παραμέτρων). Επίσης παρατηρούμε ότι η 28^η παρατήρηση παρουσιάζει μεγάλο studentized residual οπότε μπορεί να θεωρηθεί ως έκτροπη παρατήρηση. Όσον αφορά τις έκτροπες παρατηρήσεις ή τις παρατηρήσεις με μεγάλη επιρροή, βεβαιωνόμαστε ότι έχουν καταγραφεί και περαστεί σωστά στα δεδομένα. Αν δεν είμαστε απολύτως βέβαιοι για κάτι τέτοιο, μία συντηρητική απόφαση θα ήταν να εξαιρεθούν από το μοντέλο.



Για τον έλεγχο κανονικότητας των παρατηρήσεων πραγματοποιούμε ένα τεστ *Kolmogorov - Smirnov* για τα κατάλοιπα (Analyse/non-parametric tests/K-S test) από όπου λαμβάνουμε τον πίνακα:

One-Sample Kolmogorov-Smirnov Test

		Studentized Residual
N		30
Normal Parameters ^{a,b}	Mean	5,1754E-03
	Std. Deviation	1,0139350
Most Extreme Differences	Absolute	,090
	Positive	,070
	Negative	-,090
Kolmogorov-Smirnov Z		,494
Asymp. Sig. (2-tailed)		,968

a. Test distribution is Normal.

b. Calculated from data.

Το p-value είναι πολύ υψηλό (0.968) οπότε μπορούμε να δεχτούμε ότι τα κατάλοιπα προέρχονται από κανονική κατανομή (για να θεωρηθεί αξιόπιστο το τεστ υποθέτουμε ότι η συσχέτιση μεταξύ των καταλοίπων είναι αμελητέα).

viii) Ένας τρόπος να εξετάσουμε αν το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ βελτιώνει ουσιαστικά το προηγούμενο είναι να κάνουμε τον έλεγχο $\beta_3 = 0$ (επίσης, θα μπορούσαμε να δούμε αν προκύπτει σημαντική αύξηση του R^2 ή του $R^2(\text{adj})$ κ.τ.λ.). Από την εφαρμογή του μοντέλου αυτού λαμβάνουμε:

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-24,022	31,954		-,752	,459
	X1	2,403	,502	,499	4,790	,000
	X2	4,688	,517	,994	9,070	,000
	X12	-9,11E-03	,008	-,162	-1,132	,268

a. Dependent Variable: Y

($X_{12} = X_1 * X_2$) απ' όπου συμπεραίνουμε ότι η συνεισφορά δεν είναι σημαντική (p-value = 0.268 > 0.05).

2. Για να εξετάσουμε ποιο από τα παρακάτω τρία μοντέλα

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad Y = \beta_0 + \beta_2 X_2 + \varepsilon, \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

είναι το καλύτερο, με βάση το R^2 , το $R^2(\text{adj})$, και το $C_p = \frac{SSE_p}{MSE} - (n - 2p)$ μπορούμε να εκτελέσουμε την διαδικασία της παλινδρόμησης 3 φορές, μία για κάθε μοντέλο. Εναλλακτικά μπορούμε να δούμε και τα τρία μοντέλα μαζί αν εργαστούμε ως εξής:

1) Ανοίγουμε την διαδικασία Analyze/regression/linear θέτοντας Dependent: Y και independent X_1 με Method: Enter (Block 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$)

2) Στη συνέχεια, επιλέγουμε Next για να πάμε στο Block 2 όπου και θέτουμε Dependent: Y και independent X_2 με Method: Enter (προστίθεται και η X_2 στο μοντέλο του Block 1 οπότε προκύπτει το $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ στο Block 2)

3) Τέλος, επιλέγουμε και πάλι Next για να πάμε στο Block 3 όπου και θέτουμε Dependent: Y και independent X_1 με Method: Remove (αφαιρείται η X_1 από το μοντέλο του Block 2 οπότε προκύπτει το $Y = \beta_0 + \beta_2 X_2 + \varepsilon$ στο Block 3).

Πατώντας OK λαμβάνουμε τα αποτελέσματα:

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	X1 ^a		Enter
2	X2 ^a		Enter
3		X1 ^b	Remove

a. All requested variables entered.

b. All requested variables removed.

c. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,393 ^a	,154	,124	132,8935
2	,966 ^b	,934	,929	37,7795
3	,882 ^c	,778	,770	68,1224

a. Predictors: (Constant), X1

b. Predictors: (Constant), X1, X2

c. Predictors: (Constant), X2

ANOVA^d

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	90208,359	1	90208,359	5,108	,032 ^a
	Residual	494499,0	28	17660,679		
	Total	584707,4	29			
2	Regression	546170,5	2	273085,242	191,331	,000 ^b
	Residual	38536,883	27	1427,292		
	Total	584707,4	29			
3	Regression	454768,7	1	454768,671	97,996	,000 ^c
	Residual	129938,7	28	4640,668		
	Total	584707,4	29			

a. Predictors: (Constant), X1

b. Predictors: (Constant), X1, X2

c. Predictors: (Constant), X2

d. Dependent Variable: Y

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	210,502	54,806		3,841	,001
	X1	1,890	,836	,393	2,260	,032
2	(Constant)	4,828	19,369		,249	,805
	X1	1,902	,238	,395	8,002	,000
	X2	4,165	,233	,883	17,873	,000
3	(Constant)	116,895	24,128		4,845	,000
	X2	4,160	,420	,882	9,899	,000

a. Dependent Variable: Y

Άρα για τα τρία μοντέλα παίρνουμε τον παρακάτω πίνακα:

Model	R^2	$R^2(\text{adj})$	SSE_p	$C_p = \frac{SSE_p}{MSE} - (n - 2p)$
1	0.154	0.124	494499.008	$346 - 26 = 320.4$
2	0.934	0.929	38536.883	$27.0 - 24 = 3$
3	0.778	0.770	129938.696	$91.0 - 29 = 62$

(MSE (πλήρους μοντέλου) = 1427.292)

Οπότε, με βάση και τα τρία κριτήρια, το καλύτερο μοντέλο είναι το πλήρες μοντέλο Y, X_1, X_2 (μοντέλο 2). Υπενθυμίζεται ότι σύμφωνα με το R^2 , καλύτερο θεωρείται το μοντέλο στο οποίο δεν αυξάνεται σημαντικά το R^2 με την πρόσθεση νέων ανεξάρτητων μεταβλητών (μοντέλο που «σταθεροποιεί» το R^2). Σύμφωνα με το $R^2(\text{adj})$ καλύτερο είναι το μοντέλο με το μεγαλύτερο $R^2(\text{adj})$ (ισοδύναμα μπορούμε να δούμε ποιο μοντέλο έχει το μικρότερο MSE_p διότι $R^2(\text{adj}) = 1 - (n-1)MSE_p/SST$). Τέλος, σύμφωνα με τον δείκτη C_p του Mallows's καλύτερο είναι το μοντέλο με το μικρότερο C_p που βρίσκεται κοντά στο p ($C_p \approx p$).

Παρατηρούμε στους παραπάνω πίνακες ότι τα εκτιμημένα β_1, β_2 δεν αλλάζουν σημαντικά στα τρία μοντέλα:

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	1,890	
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	1,902	4,165
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$		4,160

Αυτό συμβαίνει όταν οι ανεξάρτητες (ή ερμηνευτικές) μεταβλητές X_1, X_2 είναι σχεδόν ασυσχέτιστες (δειγματικός συντελεστής συσχέτισης κοντά στο 0). Πράγματι, σε παραπάνω πίνακα είχαμε δει ότι το Pearson Correlation μεταξύ του X_1 και του X_2 είναι -0.003 . Αν οι ανεξάρτητες μεταβλητές έχουν ισχυρή συσχέτιση (π.χ. Pearson Correlation κοντά στο 1 ή στο -1) τότε παρουσιάζεται το φαινόμενο της πολυσυγγραμμικότητας (multicollinearity). Αυτό το φαινόμενο παρουσιάζεται στα δεδομένα του πίνακα Β. Συγκεκριμένα παίρνοντας τα 3 παραπάνω μοντέλα αυτή τη φορά με τα δεδομένα του πίνακα Β λαμβάνουμε:

Correlations

		Y	X1	X2
Pearson Correlation	Y	1,000	,843	,878
	X1	,843	1,000	,924
	X2	,878	,924	1,000
Sig. (1-tailed)	Y	,	,000	,000
	X1	,000	,	,000
	X2	,000	,000	,
N	Y	20	20	20
	X1	20	20	20
	X2	20	20	20

(παρατηρούμε υψηλή θετική συσχέτιση 0.924 μεταξύ του X_1 και του X_2)

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	X1 ^a	,	Enter
2	X2 ^a	,	Enter
3	, ^a	X1 ^b	Remove

- a. All requested variables entered.
b. All requested variables removed.
c. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,843 ^a	,711	,695	2,8198
2	,882 ^b	,778	,752	2,5432
3	,878 ^c	,771	,758	2,5102

- a. Predictors: (Constant), X1
b. Predictors: (Constant), X1, X2
c. Predictors: (Constant), X2

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,496	3,319		-,451	,658
	X1	,857	,129	,843	6,656	,000
2	(Constant)	-19,174	8,361		-2,293	,035
	X1	,222	,303	,219	,733	,474
	X2	,659	,291	,676	2,265	,037
3	(Constant)	-23,634	5,657		-4,178	,001
	X2	,857	,110	,878	7,786	,000

a. Dependent Variable: Y

Παρατηρούμε στους παραπάνω πίνακες ότι τα εκτιμημένα β_1, β_2 αλλάζουν σημαντικά στα τρία μοντέλα:

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	0,857	
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	0,222	0,659
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$		0.857

Μάλιστα παρατηρούμε ότι ενώ στο πρώτο μοντέλο $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ η X_1 είναι σημαντική, όταν στο μοντέλο προστίθεται η X_2 ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$) η X_1 γίνεται μη-σημαντική ενώ τη θέση της παίρ-

νει η X_2 . Μάλιστα, καλύτερο μοντέλο σύμφωνα με το $R^2(\text{adj})$ φαίνεται να είναι το $Y = \beta_0 + \beta_2 X_2 + \varepsilon$. Εδώ μπορούμε π.χ. να δώσουμε την εξής ερμηνεία: Η X_2 είναι πράγματι αυτή που επηρεάζει την Y αλλά όταν εφαρμόζουμε το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, η X_1 φαίνεται σημαντική διότι είναι πολύ «κοντά» στην X_2 (το t-test για το β_1 στο μοντέλο 1 στην ουσία ελέγχει αν η είσοδος του X_1 στο μοντέλο $Y = \beta_0 + \varepsilon$ είναι σημαντική). Αν όμως υπάρχει και η X_2 στο μοντέλο τότε το αντίστοιχο t-test για το β_1 στο μοντέλο 2 ελέγχει αν η είσοδος της X_1 στο μοντέλο $Y = \beta_0 + \beta_2 X_2 + \varepsilon$ είναι σημαντική. Είναι φυσικό να μην φαίνεται τώρα σημαντική η X_1 διότι δεν συνεισφέρει τίποτε νέο στο μοντέλο αφού υπάρχει ήδη η X_2 η οποία ερμηνεύει ικανοποιητικά τη μεταβλητότητα του Y .

Στις περιπτώσεις όπου παρουσιάζεται υψηλή πολυσυγγραμμικότητα (μπορεί να εντοπισθεί και μέσω του δείκτη VIF (variance inflation factor) π.χ. όταν $\text{VIF} > 10$: ο δείκτης VIF_i που δίνεται στο πακέτο είναι ίσος με $(1 - R_i^2)^{-1}$ όπου R_i^2 είναι ο συντελεστής προσδιορισμού όταν κάνουμε παλινδρόμηση της X_i (ως dependent) με τις υπόλοιπες X (ως independent)), εντοπίζουμε τις ομάδες των ερμηνευτικών μεταβλητών με υψηλή συσχέτιση και συνήθως παραμένει μόνο μία μεταβλητή από κάθε ομάδα. Σε αρκετές περιπτώσεις αυτό δεν είναι εύκολο οπότε είναι ανάγκη να χρησιμοποιήσουμε άλλες μεθόδους (π.χ. principal components regression ή ridge regression). Σημειώνεται ότι η ύπαρξη πολυσυγγραμμικότητας μπορεί να οδηγήσει και σε μία σχεδόν μηδενική ορίζουσα του $\mathbf{X}'\mathbf{X}$, με αποτέλεσμα να υπάρχει κίνδυνος εμφάνισης σοβαρών σφαλμάτων στρογγύλευσης στην διαδικασία υπολογισμού των εκτιμήσεων των β_i . Για το λόγο αυτό προτείνεται η κανονικοποίηση των Y, X_1, X_2 πριν την εφαρμογή της παλινδρόμησης στο πακέτο ώστε να παίρνουν τιμές στο $(-1, 1)$ με σκοπό να μειωθούν τα σφάλματα στρογγύλευσης (συνήθως η κανονικοποίηση αυτή γίνεται αυτόματα από το πακέτο).