

ΙΙΜΣ ΕΠΑΓΓΕΛΜΑΤΙΚΗ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΙΚΗ ΥΓΕΙΑ, ΔΙΑΧΕΙΡΙΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΚΗ ΑΡΩΣΤΙΑΣ
ΑΚ. ΕΤΟΣ 2006-2007, 3ο εξάμηνο

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

ΜΑΘΗΜΑ 11

Συμπερασματολογία για την επίδραση πολλών μεταβλητών σε μια ποσοτική (Πολλαπλή Παλινδρόμηση)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-1

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική ΠΕΡΙΕΧΟΜΕΝΑ

- Πολλαπλή Παλινδρόμηση (multiple regression)
 - ✓ Γενίκευση του μοντέλου
 - ✓ Ερμηνεία παραμέτρων
 - ✓ Παράδειγμα 11-1 και εφαρμογή στο SPSS
 - ✓ Διαδικασίες επιλογής μοντέλων και μεταβλητών (και Εφαρμογή στο SPSS – παράδειγμα 11-1)
 - ✓ Το πρόβλημα της πολυσυγγραμμικότητας (και Εφαρμογή στο SPSS – παράδειγμα 11-1, Παράδειγμα 11-2: προσομοιωμένα δεδομένα)
 - ✓ ΠΑΡΑΔΕΙΓΜΑ 11-3 WORLD95
- Δείκτες μερικής συσχέτισης (Partial correlations) (+ Συνέχεια παραδείγματος 11-1)
- Πολυωνυμική Παλινδρόμηση
- Χρήση κατηγορικών συμμεταβλητών σε παλινδρομικά μοντέλα
- Σχέση παλινδρόμησης και ANOVA/t-tests
- Αναλ. Συνδιακύμανσης (ANCOVA)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-2

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.1. Γενίκευση του μοντέλου

Έστω ότι έχουμε $p+1$ ποσοτικές μεταβλητές

- Y : απόκρισης ή εξαρτημένη
- X_1, X_2, \dots, X_p : επεξηγηματικές ή ανεξάρτητες μεταβλητές

Το μοντέλο:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ ή ισοδύναμα
- $Y \sim N(\mu, \sigma^2)$, $E(Y) = \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$

Μοντέλο και δεδομένα:

- Y_i, X_i ζεύγη τιμών για $i=1, 2, \dots, n$
 - $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$
 - $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-3

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.1. Γενίκευση του μοντέλου

ΠΡΟΫΠΟΘΕΣΕΙΣ: ΙΔΙΕΣ

ΕΡΜΗΝΕΙΑ: ΛΙΓΟ ΔΙΑΦΟΡΕΤΙΚΗ

EXTRA ΠΡΟΒΛΗΜΑ: ΠΟΛΥ-ΣΥΓΓΡΑΜΙΚΟΤΗΤΑ μεταξύ επεξηγηματικών μεταβλητών (ύπαρξη γραμμικών σχέσεων μεταξύ επεξηγηματικών μεταβλητών) [Multi-collinearity]

EXTRA ΕΡΩΤΗΜΑ: ΕΠΙΛΟΓΗ ΣΗΜΑΝΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-4

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.2. Ερμηνεία παραμέτρων

- β_0 :** Αναμενόμενη τιμή της Y όταν όλες οι X είναι μηδέν
 - Αν πάρουμε ως επεξηγηματικές τις μεταβλητές κεντραρισμένες ως προς τον μέσο
 $X_j^* = X_j - \bar{X}_j$
 - Τότε αντιστοιχεί στην αναμενόμενη τιμή του Y όταν όλες οι X είναι ίσες με τους δειγματικούς μέσους (ένα τυπικό – μέσο άτομο του δείγματος)
- β_j :** Αναμενόμενη μεταβολή τιμής της Y όταν η X_j αυξηθεί κατά μία μονάδα αλλά οι υπόλοιπες X παραμείνουν σταθερές.
 - Επίδραση (effect) της μεταβλητής X_j στην Y διορθωμένη (adjusted) για τις επιδράσεις των υπόλοιπων μεταβλητών
 - Επίδραση διορθωμένη ως προς τις X μεταβλητές σημαίνει ότι τις έχουμε λάβει υπόψη μας στην ανάλυση μας

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-5

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

Παράδειγμα 11-1

Ένα κτηματομεσίτης έκανε έρευνα αγοράς σε 3 περιοχές με σκοπό να φτιάξει ένα απλό μοντέλο για να υπολογίζει – προβλέπει τις αναμενόμενες τιμές των σπιτιών ανά περιοχή.

Έτσι πήρε 10 σπίτια σε κάθε περιοχή τυχαία επιλεγμένα και μέτρησε

- ✓ Τιμή σε δολάρια
- ✓ Χώρος Κατοικίας (κτηρίου)
- ✓ Έκταση οικοπέδου
- ✓ Περιοχή που βρίσκεται

Sampled house	Price	Living area sq. ft.	Lot size sq. ft.	Subdivision
1	102,200	1,500	12,000	A
2	103,950	1,200	10,000	A
3	87,900	1,200	10,000	A
4	110,000	1,600	15,000	A
5	87,000	1,400	12,000	A
6	95,700	1,200	10,000	A
7	113,600	1,600	15,000	A
8	109,600	1,500	12,000	A
9	110,800	1,500	12,000	A
10	90,600	1,300	12,000	A
11	109,000	1,600	13,000	B
12	133,000	1,900	15,000	B
13	134,000	1,800	15,000	B
14	120,300	2,000	17,000	B
15	137,000	2,000	17,000	B
16	122,400	1,700	15,000	B
17	121,700	1,800	15,000	B
18	126,000	1,900	16,000	B
19	128,000	2,000	16,000	B
20	117,500	1,600	13,000	B
21	158,700	2,400	18,000	C
22	186,800	2,600	18,000	C
23	172,400	2,300	16,000	C
24	151,200	2,200	16,000	C
25	178,100	2,800	20,000	C
26	182,300	2,700	20,000	C
27	195,850	3,000	22,000	C
28	168,000	2,400	18,000	C
29	199,400	2,500	20,000	C
30	163,000	2,400	18,000	C

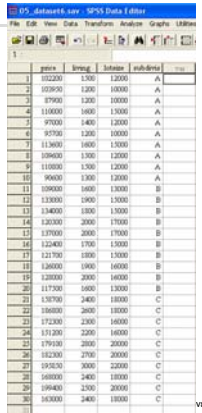
ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-6

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

Παράδειγμα 11-1

- Μονάδα μελέτης: Κατοικία
- n=30
- Μεταβλητές p=4
 - ✓ **Price:** Τιμή σε δολάρια
 - ✓ **Living:** Χώρος Κατοικίας (κτηρίου)
 - ✓ **Lotsize:** Έκταση οικοπέδου
 - ✓ **Subdivis:** Περιοχή που βρίσκεται το σπίτι



ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-7

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΑΝΑΛΥΣΗ – ΒΗΜΑΤΑ

- Ανάλυση ανά μία μεταβλητή
- Διαγραμματική απεικόνιση (Scatter-plots)
- Δείκτες συσχέτισης
- Μοντέλο Παλινδρόμησης
- Επιλογή Μεταβλητών
- Έλεγχος πολυσυγγραμμικότητας
- Έλεγχος Προϋποθέσεων (Ανάλυση καταλοίπων)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-8



11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΔΕΙΚΤΕΣ ΓΡΑΜΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ

Correlations

		price Price in US Dollars	living Size of Living Area in Sq. ft.	lotsize Lot Size in Sq. ft.
price Price in US Dollars	Pearson Correlation	1	.963**	.907**
	Sig. (2-tailed)		.000	.000
	N	30	30	30
living Size of Living Area in Sq. ft.	Pearson Correlation	.963**	1	.961**
	Sig. (2-tailed)	.000		.000
	N	30	30	30
lotsize Lot Size in Sq. ft.	Pearson Correlation	.907**	.961**	1
	Sig. (2-tailed)	.000	.000	
	N	30	30	30

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

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-11

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Επιλογή μενού στο SPSS

Εξαρτημένη μεταβλητή

ανεξάρτητες μεταβλητές

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Περίληπτικός πίνακας

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.965 ^a	.930	.925	9082.657

a. Predictors: (Constant), lotsize Lot Size in Sq.ft., living Size of Living Area in Sq. ft

$\sigma = 9082.66\$$

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-13

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Πίνακας ανάλυσης Διακύμανσης

Ελέγχει την $H_0: \beta_1 = \beta_2 = 0$
δηλαδή αν υπάρχει σημαντική
διαφοροποίηση από το
σταθερό μοντέλο

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29763652086.191	2	1.488E+1010	180.397	.000 ^b
	Residual	2227355913.810	27	82494663.47		
	Total	31991008000.000	29			

a. Predictors: (Constant), lotsize Lot Size in Sq.ft., living Size of Living Area in Sq. ft
b. Dependent Variable: price Price in US Dollars

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-14

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Πίνακας εκτιμήσεων των παραμέτρων

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22149.972	9552.441		2.319	.028
	living Size of Living Area in Sq. ft	77.024	11.967	1.180	6.436	.000
	lotsize Lot Size in Sq.ft.	-2.345	1.900	-.226	-1.234	.228

a. Dependent Variable: price Price in US Dollars

Τιμή = 22150 + 77 Μέγ.Σπιτ (sq.ft.) - 2.35 Μέγ.Οικ.(sq.ft.)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-15

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Ερμηνεία παραμέτρων

Τιμή = 22150 + 77 Μέγ.Σπιτ (sq.ft.) - 2.35 Μέγ.Οικ.(sq.ft.)

- Όταν το οικόπεδο και το κτίριο έχουν μηδενική έκταση (???) τότε η αναμενόμενη τιμή είναι 22150\$
 - ✓ Δεν στέκει ως ερμηνεία
 - ✓ Μόνο μπορούμε να θεωρήσουμε το ποσό των 22150 κάποια πάγια έξοδα
 - ✓ Γενικά θα είναι μάλλον πιο λογικό να αφαιρέσουμε τη σταθερά (αν και είναι στατιστικά σημαντική)
- Αύξηση της έκτασης του κτιρίου κατά 1 sq.ft. συνεπάγεται αύξηση κατά 77\$ όταν η έκταση του οικοπέδου παραμένει σταθερή
- ΠΙΟ ΣΩΣΤΗ ΕΡΜΗΝΕΙΑ:** Αν συγκρίνουμε δύο συμβόλαια με ίδια έκταση οικοπέδου και διαφορά στο μέγεθος του κτιρίου κατά 1 sq.ft. τότε αναμένουμε διαφορά ίση με 77\$
- ΠΙΟ ΣΩΣΤΗ ΕΡΜΗΝΕΙΑ:** Αν συγκρίνουμε δύο συμβόλαια με ίδια έκταση κτιρίου και διαφορά στο μέγεθος του οικοπέδου κατά 1 sq.ft. τότε αναμένουμε διαφορά ίση με 2,35\$ (υπερ του μεγαλύτερου οικοπέδου)???? [ΔΕΝ ΣΤΕΚΕΙ - ΟΜΩΣ ΕΙΝΑΙ ΣΤΑΤΙΣΤΙΚΑ ΜΗ ΣΗΜΑΝΤΙΚΟ]

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-16

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Αφαίρεση σταθεράς

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-17

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Αφαίρεση σταθεράς

Model Summary

Model	R	R Square ^a	Adjusted R Square	Std. Error of the Estimate
1	.998 ^b	.995	.995	9766.756

a. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

b. Predictors: lotsize Lot Size in Sq.ft., living Size of Living Area in Sq. ft

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-18

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Αφαίρεση σταθεράς

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	living Size of Living Area in Sq. ft	62.808	11.051	.902	5.683	.000
	lotsize Lot Size in Sq.ft.	.850	1.407	.096	.604	.551

a. Dependent Variable: price Price in US Dollars
b. Linear Regression through the Origin

Τιμή = 62.8 Μεγ.Σηπτ (sq.ft) + 0.85 Μέγ.Οικ.(sq.ft.)

Στατιστικά σημαντικό (διάφορο του μηδενός)
Δεν είναι Στατιστικά σημαντικό (άρα ίσο με μηδέν)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-19

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Ερμηνεία παραμέτρων (Χωρίς σταθερά)

Τιμή = 62.8 Μεγ.Σηπτ (sq.ft) + 0.85 Μέγ.Οικ.(sq.ft.)

- Όταν το οικοπέδο και το κτίριο έχουν μηδενική έκταση τότε η αναμενόμενη τιμή είναι 0\$ (πολύ λογικότερο ως προσέγγιση)
- Αύξηση της έκτασης του κτιρίου κατά 1 sq.ft. συνεπάγεται αύξηση κατά 62.8\$ όταν η έκταση του οικοπέδου παραμένει σταθερή
- ΠΙΟ ΣΩΣΤΗ ΕΡΜΗΝΕΙΑ: Αν συγκρίνουμε δύο συμβόλαια με ίδια έκταση οικοπέδου και διαφορά στο μέγεθος του κτιρίου κατά 1 sq.ft. τότε αναμένουμε διαφορά ίση με 62.8\$
- ΠΙΟ ΣΩΣΤΗ ΕΡΜΗΝΕΙΑ: Αν συγκρίνουμε δύο συμβόλαια με ίδια έκταση κτιρίου και διαφορά στο μέγεθος του οικοπέδου κατά 1 sq.ft. τότε αναμένουμε αύξηση της τιμής κατά 0.85\$ [ΕΙΝΑΙ ΣΤΑΤΙΣΤΙΚΑ ΣΗΜΑΝΤΙΚΟ]

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-20

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Μοντέλο με σταθερά και Μεταβλητές κεντραρισμένες στο μέσο

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.965 ^a	.930	.925	9082.657

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134230.0	1658.259		80.946	.000
	liv.c	77.024	11.967	1.180	6.436	.000
	lot.c	-2.345	1.900	-.226	-1.234	.228

a. Dependent Variable: price Price in US Dollars

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνειο 11-21

Ενα μέσο σπίτι (εκτ.1920 sq.ft.) κοστίζει περίπου 134230\$

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Έλεγχοι Προϋποθέσεων

- Κανονικά θα έπρεπε να αφαιρέσουμε και το LOTSIZE. Ας πούμε ότι το κρατάμε για να έχουμε ένα πιο ρεαλιστικό μοντέλο
- ΚΑΝΟΝΙΚΟΤΗΤΑ (QQPLOT, HISTOGRAM, SW/KS TESTS)
- ΟΜΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑ – ΓΡΑΜΜΙΚΟΤΗΤΑ – ΑΚΡΑΙΕΣ ΤΙΜΕΣ (PRED-ST.RESIDUALS)
- ΤΥΧΑΙΟΤΗΤΑ (LINE PLOT, RUNS TEST, DW TEST, ACF+PACF)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνειο 11-23

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Μοντέλο χωρίς σταθερά και Μεταβλητές κεντραρισμένες στο μέσο

Model Summary

Model	R	R Square ^a	Adjusted R Square	Std. Error of the Estimate
1	.228 ^b	.052	-.016	139227.214

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	liv.c	77.027	183.438	.279	.420	.678
	lot.c	-2.346	29.123	-.054	-.081	.936

a. Dependent Variable: price Price in US Dollars

b. Linear Regression through the Origin

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II © 11-22

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Έλεγχοι Προϋποθέσεων – Κανονικότητα

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Έλεγχοι Προϋποθέσεων – Κανονικότητα

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RES_1 Unstandardized Residual	.103	30	.200*	.973	30	.627

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-25

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Έλεγχοι Προϋποθέσεων
Ομοσκεδαστικότητα, Γραμμικότητα, Ακραίες τιμές

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-26

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Έλεγχοι Προϋποθέσεων
Ομοσκεδαστικότητα, Γραμμικότητα, Ακραίες τιμές

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-27

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.3. Παράδειγμα 11-1 και εφαρμογή στο SPSS

ΜΟΝΤΕΛΟ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
Έλεγχοι Προϋποθέσεων
Ομοσκεδαστικότητα, Γραμμικότητα, Ακραίες τιμές

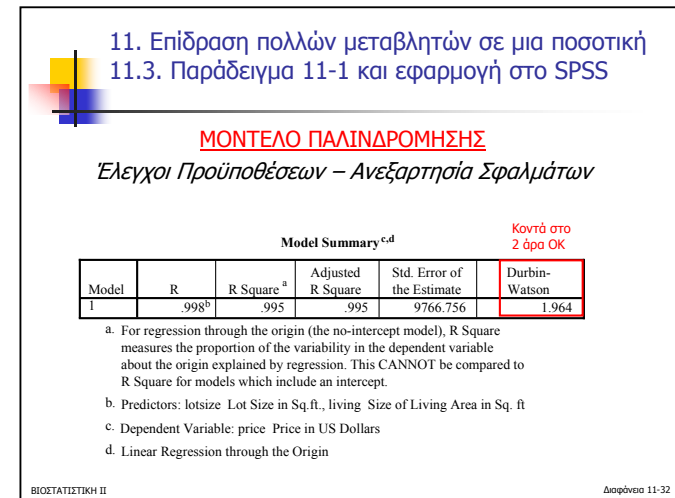
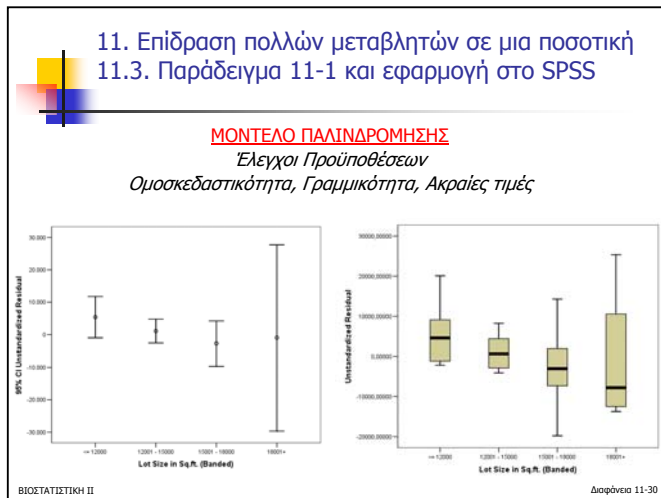
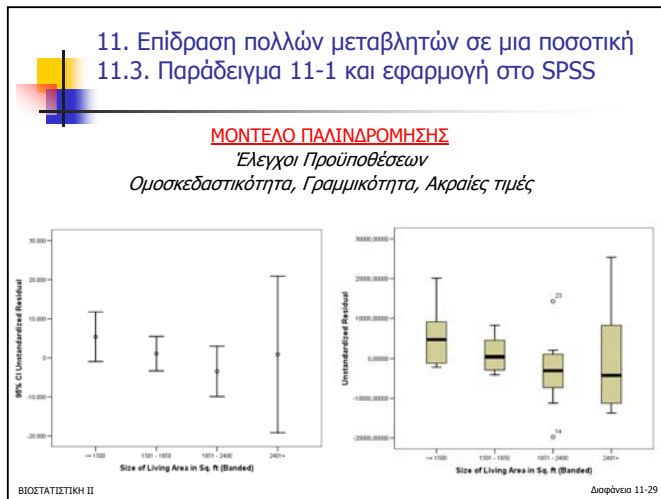
Test of Homogeneity of Variances

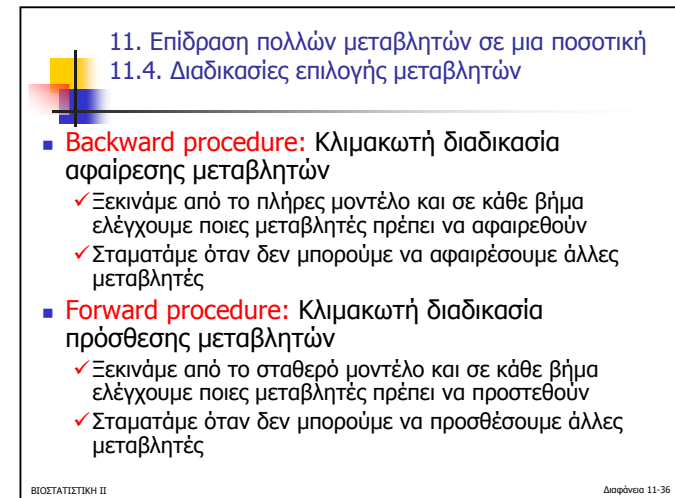
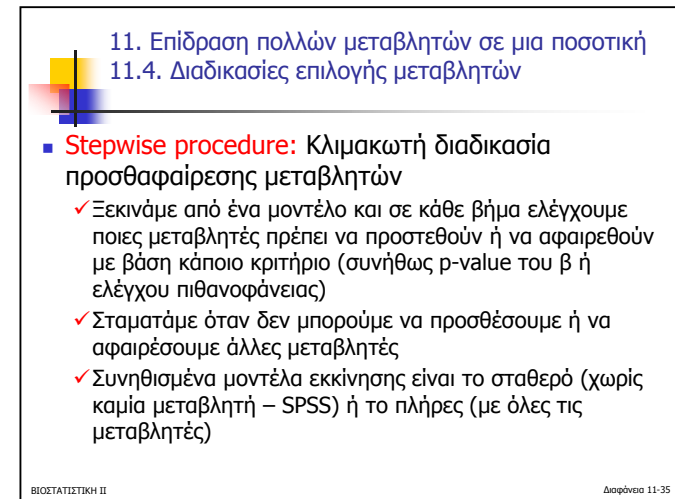
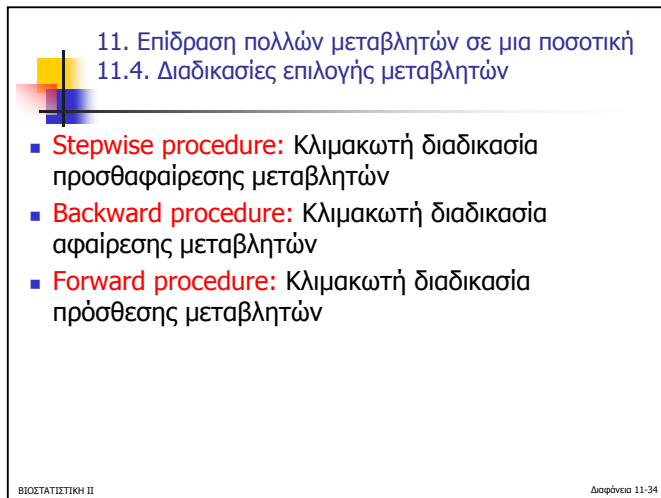
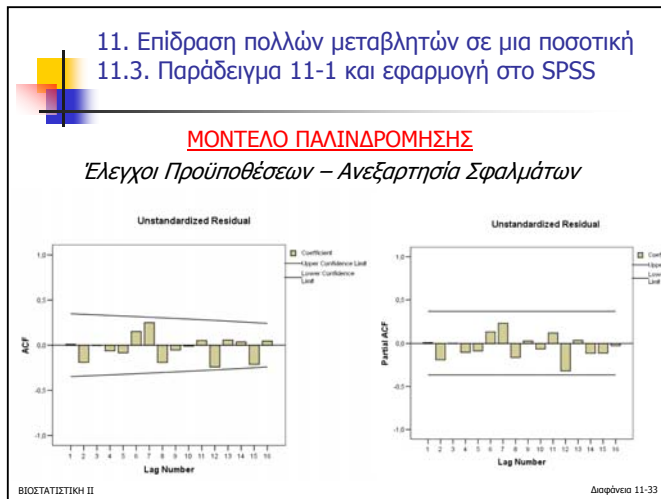
RES_1 Unstandardized Residual

Quartiles of LIVING	Levene Statistic	df1	df2	Sig.
	2.881	3	26	.055

Quartiles of LOTSIZE	Levene Statistic	df1	df2	Sig.
	2.735	3	26	.064

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-28





11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.4. Διαδικασίες επιλογής μεταβλητών

- **ΚΡΙΤΗΡΙΑ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ – ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ**
 - ✓ F – test από τη διαφορά διαδοχικών μοντέλων
 - ✓ t-test για β
 - ✓ BIC, AIC, Cp (Sp/RS)
- **ΛΕΠΤΟΜΕΡΕΙΕΣ**
 - ✓ Καλύτερη η Stepwise γιατί κάνει διπλούς ελέγχους
 - ✓ Σημείο εκκίνησης το πλήρες.
 - ✓ Αν έχουμε πολλές μεταβλητές τότε σημείο εκκίνησης το σταθερό
 - ✓ Διαφορετικές διαδικασίες μπορούν να καταλήξουν σε άλλα μοντέλα
 - ✓ Διαφορετικά σημεία εκκίνησης μπορεί να καταλήξουν σε διαφορετικό μοντέλο.

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-37

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.4. Διαδικασίες επιλογής μοντέλων και μεταβλητών
 (Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-38

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.4. Διαδικασίες επιλογής μοντέλων και μεταβλητών
 (Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

Variables Entered/Removed ^{a,c}				Model Summaries ^d				
Model	Variables Entered	Variables Removed	Method	Model R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	lotsize Lot Size in Sq. ft., living Size of Living Area in Sq. ft.		Enter	.998 ^b	.995	.995	9766.756	
2	lotsize Lot Size in Sq. ft.		Backward (criterion: Probability of F-to-remove >= .100)	.998 ^c	.995	.995	9659.215	1.961

a. All requested variables entered.
 b. Dependent Variable: price Price in US Dollars
 c. Linear Regression through the Origin
 d. For regression through the origin (the no-intercept model) the proportion of the variability in the dependent variable about the origin explained by the regression. This CAN BE R Square for models which include an intercept.
 e. Predictors: lotsize Lot Size in Sq.ft., living Size of Living Area in Sq. ft.
 f. Predictors: living Size of Living Area in Sq. ft.
 g. Dependent Variable: price Price in US Dollars
 h. Linear Regression through the Origin

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-39

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.4. Διαδικασίες επιλογής μοντέλων και μεταβλητών
 (Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

Coefficients ^{a,b}						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	living Size of Living Area in Sq. ft.	62.808	11.051	.902	5.683	.000
1	lotsize Lot Size in Sq.ft.	.850	1.407	.096	.604	.551
2	living Size of Living Area in Sq. ft.	69.461	.889	.998	78.149	.000

a. Dependent Variable: price Price in US Dollars
 b. Linear Regression through the Origin

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-40

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.4. Διαδικασίες επιλογής μοντέλων και μεταβλητών (Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

Excluded Variables^{a,c}

Model		Beta	In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
2	lotsize Lot Size in Sq.ft.	.096 ^b		.604	.551	.113	.007

a. Predictors in the Model: living Size of Living Area in Sq. ft

b. Dependent Variable: price Price in US Dollars

c. Linear Regression through the Origin

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-41

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.5. Το πρόβλημα της πολυσυγγραμμικότητας

- Πολυσυγγραμμικότητα (multi-collinearity)
Η υψηλή (στατιστικά) γραμμική σχέση μεταξύ μιας επεξηγηματικής μεταβλητής με τις υπόλοιπες
- Συγγραμμικότητα (collinearity)
Η τέλεια γραμμική σχέση μεταξύ μιας επεξηγηματικής μεταβλητής με τις υπόλοιπες
- Στη βιβλιογραφία πολλές φορές οι 2 όροι ταυτίζονται
- Για λεπτομέρειες βλ. Ryan (1997, σελ. 131)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-42

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.5. Το πρόβλημα της πολυσυγγραμμικότητας

■ ΠΑΡΕΝΕΡΓΕΙΣ

- Όταν υπάρχει πλήρη γραμμική σχέση ⇔ δεν μπορούν να βρεθούν εκτιμητές Μεγ.Πιθ. (ή ελ.τετρ.)
- Υψηλά τυπικά σφάλματα
- Αστάθεια εκτιμητών
- Αλλοίωση επιδράσεων (ακόμα και αλλαγή πρόσημων στις επιδράσεις)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-43

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική

11.5. Το πρόβλημα της πολυσυγγραμμικότητας

Γιατί είναι πρόβλημα;

ΛΟΓΙΚΗ ΕΠΕΞΗΓΗΣΗ

- Αν 2 μεταβλητές σχετίζονται ισχυρά μεταξύ τους, τότε μεταφέρουν παρόμοια πληροφορία (εφόσον γνωρίζοντας τη μια μπορούμε με ακρίβεια να προβλέψουμε την άλλη). Συνεπώς τέτοιες μεταβλητές δεν προσθέτουν πληροφορία όταν τις προσθέτουμε στο μοντέλο
- Παρόμοια είναι η περίπτωση αν έχουμε εξάρτηση με περισσότερες από μία μεταβλητές

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-44

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

Γιατί είναι πρόβλημα;

ΕΡΜΗΝΕΥΤΙΚΗ – ΑΡΙΘΜΗΤΙΚΗ ΕΞΗΓΗΣΗ

Έστω το παλινδρομικό μοντέλο

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

Όμως $X_2 = a + b X_1$ (τέλεια γραμμική σχέση)

Δεν μπορούμε να χρησιμοποιήσουμε την προηγούμενη ερμηνεία διότι μεταβολή στη X_1 συνεπάγεται μεταβολή και στην X_2

ΕΠΙΠΛΕΟΝ

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 (a + b X_1) + \varepsilon \\ &= (\beta_0 + a \beta_2) + (\beta_1 + \beta_2 b) X_1 + \varepsilon \end{aligned}$$

Ποια είναι η σωστή επίδραση της X_1 ;

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-45

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

Γιατί είναι πρόβλημα;

ΜΑΘΗΜΑΤΙΚΗ ΕΞΗΓΗΣΗ

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T$ είναι το διάνυσμα των εκτιμητών μέγιστης πιθανοφάνειας διάστασης $(p+1) \times 1$
- X είναι ο πίνακας σχεδιασμού ή δεδομένων διάστασης $n \times (p+1)$. Η πρώτη στήλη αναφέρεται στο σταθερό όρο και έχει όλα της τα στοιχεία της ίσα με ένα (1). Οι υπόλοιπες έχουν τα δεδομένα κάθε μεταβλητής
- y είναι το διάνυσμα διάστασης $n \times 1$ με τα δεδομένα της μεταβλητής απόκρισης.

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-46

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

Γιατί είναι πρόβλημα;

ΜΑΘΗΜΑΤΙΚΗ ΕΞΗΓΗΣΗ

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- **ΠΡΟΒΛΗΜΑ:** Αν μια μεταβλητή (δηλ. Στήλη του X) είναι γραμμικός συνδυασμός των υπόλοιπων τότε δεν υπάρχει ο αντίστροφος $(X^T X)^{-1}$
- **ΣΤΗΝ ΠΡΑΞΗ:** Σπάνια έχουμε τέλεια γραμμική σχέση. Αν όμως μια μεταβλητή σχετίζεται υψηλά με τις υπόλοιπες (δηλ. Κάνουμε παλινδρόμηση μεταξύ τους και προκύψει μεγάλο R^2) τότε έχουμε ασταθείς (unstable) εκτιμήσεις και μεγάλα τυπικά σφάλματα.

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-47

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

1. **Συσχετίσεις Pearson**
2. **Συντελεστές Πληθωρισμού Διακυμάνσεων** (variance inflation factors)
3. **Έλεγχος με ιδιοτιμές και ιδιοδιανύσματα** της μήτρας $X^T X$
4. **Αναλογίες Αποσύνθεσης διακύμανσης** (variance-decomposition proportions ή απλά variance proportions)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-48

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

- 1. Συσχετίσεις Pearson** [Δείχνουν υψηλές γραμμικές σχέσεις ανά 2 αλλά όχι για περισσότερες μεταβλητές όπως για $X_1=X_2+X_3+X_4$]
- 2. Συντελεστές Πληθωρισμού Διακυμάνσεων** (variance inflation factors)
 - ✓ $VIF(j) = (1-R_j^2)^{-1}$
 - ✓ R_j^2 = Συντελεστής προσδιορισμού που προκύπτει από την παλινδρόμηση των υπόλοιπων επεξηγηματικών μεταβλητών στην X_j .
 - ✓ $Tolerance_j = (1-R_j^2) = 1/VIF(j)$: Δείκτης Ανεκτικότητας. Δείχνει το ποσοστό της διακύμανσης που δεν εξηγείται από τις υπόλοιπες συμμεταβλητές. Χαμηλές τιμές υποδεικνύουν πρόβλημα.
 - ✓ Αν $VIF(j) > 10$ έχουμε πρόβλημα
 - ✓ Για $p=2$ (2 επεξηγηματικές μεταβλητές) τότε $VIF(i) > 10 \Leftrightarrow |r_{x_1x_2}| > 0.949$

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-49

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

- 3. Έλεγχος με ιδιοτιμές και ιδιοδιανύσματα** της μήτρας $X'X$
 - ✓ Ιδιοτιμές κοντά στο μηδέν υποδεικνύουν πρόβλημα.
 - ✓ $Condition\ Index = \sqrt{\text{Τετραγωνική ρίζα (MAX(Ιδιοτιμών)/Ιδιοτιμή)}}$
 - ✓ Αν $CI_j > 30 \Leftrightarrow$ σοβαρό πρόβλημα
 - ✓ Αν $CI_j > 15 \Leftrightarrow$ πιθανό πρόβλημα
 - ✓ Μεταβλητές που έχουν υψηλές τιμές ιδιοδιανυσμάτων είναι μεταβλητές που συμμετέχουν στην γραμμική σχέση.
- 4. Αναλογίες Αποσύνθεσης διακύμανσης** (variance-decomposition proportions ή απλά variance proportions)
Αναλογία (Ποσοστό) του VIF που προκύπτει από τη γραμμική σχέση που απεικονίζει η αντίστοιχη ιδιοτιμή (και ιδιοδιάνυσμα).

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-50

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας

ΤΡΟΠΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ

- 1. Προσεκτικός σχεδιασμός πειράματος.**
 - ✓ Όχι τυχαία X αλλά με βάση πειραματικό σχεδιασμό
 - ✓ δύσκολο στην πράξη
- 2. Αφαίρεση Προβληματικών μεταβλητών.**
 - ✓ Με μεγάλα $VIF > 10$ &
 - ✓ Αντιστοιχεί σε μικρή ιδιοτιμή
 - ✓ Αφαιρούμε μόνο μία από αυτές με μεγάλα proportion variance που αντιστοιχούν στην ίδια μικρή ιδιοτιμή
 - ✓ Τσεκάρουμε R^2 το οποίο πρέπει να αλλάξει ελάχιστα (εδώ βοηθάει και οι κλιμακωτές διαδικασίες επιλογής μεταβλητών)
 - ✓ Προσπαθούμε να έχουμε $CI < 15$ (ή έστω $CI < 30$)
- 3. Χρήση ορθογώνιου μετασχηματισμού (Κυρίες συνιστώσες) των X .**
 - ✓ Δύσκολη ερμηνεία

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-51

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

The image shows two overlapping SPSS dialog boxes. The background box is the 'Linear Regression' dialog, with 'price' selected as the dependent variable and 'living', 'lotsize', and 'indbns' as independent variables. The 'Statistics' button is highlighted with a blue box. The foreground box is the 'Linear Regression: Statistics' sub-dialog, where the 'Collinearity diagnostics' checkbox is checked and highlighted with a red box. A blue arrow points from the 'Statistics' button in the main dialog to the 'Collinearity diagnostics' checkbox in the sub-dialog.

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II

Διαφάνεια 11-52

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

Coefficients ^{a,b}								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	living Size of Living Area in Sq. ft.	62.808	11.051	.902	5.683	.000	.007	151.204
	lotsize Lot Size in Sq. ft.	850	1.407	.096	.604	.551	.007	151.204

a. Dependent Variable: price Price in US Dollars
b. Linear Regression through the Origin

VIF>10 ⇔ ΠΡΟΒΛΗΜΑ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-53

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(Εφαρμογή στο SPSS: Συνέχεια παραδείγματος 11-1)

Collinearity Diagnostics ^{a,b}					
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions	
				living Size of Living Area in Sq. ft.	lotsize Lot Size in Sq. ft.
1	1	1.997	1.000	.00	.00
	2	.003	24.552	1.00	1.00

a. Dependent Variable: price Price in US Dollars
b. Linear Regression through the Origin

Στο γραμμικό συνδυασμό της προβληματικής μεταβλητής συμμετέχουν οι 2 μεταβλητές

ΠΡΟΒΛΗΜΑΤΙΚΗ ΙΔΙΟΤΗΤΗ ΜΕ CP>15

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-54

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ ΠΡΟΣΟΜΙΩΜΕΝΑ ΔΕΔΟΜΕΝΑ

- n=100
- $X_2, X_3, X_4, X_5, X_6, X_7 \sim N(0,1)$
- $X_1 = X_2 + X_3 + X_4$
- $Y = 4 + X_2 - 3 X_4 + 5 X_6 + \epsilon, \quad \epsilon \sim N(0, 0.25 = 0.5^2)$

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-55

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-56

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

```

    COMPUTE x2 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE x3 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE x4 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE x5 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE x6 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE x7 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE x8 = RV.NORMAL(0,1) .
    EXECUTE .
    COMPUTE e = RV.NORMAL(0,0.5) .
    EXECUTE .
    
```

COPY+PASTE

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-57

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-58

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

```

    COMPUTE x1 = x2+x3+x4 .
    EXECUTE .
    COMPUTE y = 4 + x2 - 3*x4 +5* x6 + e .
    EXECUTE .
    
```

ΚΑΝΟΥΜΕ ΤΗΝ ΠΑΛΙΝΔΡΟΜΗΣΗ ΤΩΝ Χ ΣΤΗΝ Υ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-59

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

**ΤΟ ΠΡΟΓΡΑΜΜΑ ΑΠΟΦΑΣΙΣΕ ΑΠΟ ΜΟΝΟ ΤΟΥ ΝΑ
 ΑΦΑΙΡΕΣΕΙ ΜΙΑ ΑΠΟ ΤΙΣ ΜΕΤΑΒΛΗΤΕΣ ΠΟΥ
 ΕΜΠΛΕΚΟΝΤΑΙ ΣΤΗΝ ΤΕΛΕΙΑ ΓΡΑΜΜΙΚΗ**

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta				Tolerance	VIF
1	(Constant)	4.030	.044			91.126	.000		
	x2	.943	.059	.168	.15970	.000	.500	2.000	
	x4	-.034	.060	-.484	-.50909	.000	.608	1.645	
	x5	-.028	.051	-.004	-.536	.993	.928	1.078	
	x6	4.930	.043	.867	.114906	.000	.967	1.034	
	x7	.046	.050	.007	.903	.369	.908	1.101	
	x8	-.025	.048	-.004	-.514	.609	.844	1.185	
	x1	.022	.042	.006	.520	.604	.416	2.404	

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	x3	.*				.000		.000

^a Predictors in the Model: (Constant), x1, x6, x5, x7, x8, x4, x2
^b Dependent Variable: y

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-60

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΤΟ ΠΡΟΓΡΑΜΜΑ ΑΠΟΦΑΣΙΣΕ ΑΠΟ ΜΟΝΟ ΤΟΥ ΝΑ ΑΦΑΙΡΕΣΕΙ ΜΙΑ ΑΠΟ ΤΙΣ ΜΕΤΑΒΛΗΤΕΣ ΠΟΥ ΕΜΠΛΕΚΟΝΤΑΙ ΣΤΗΝ ΤΕΛΕΙΑ ΓΡΑΜΜΙΚΗ

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	x2	x4	x5	x6	x7	x8	x1
1	1	1.926	1.000	.00	.06	.06	.01	.00	.02	.04	.09
2	2	1.424	1.163	.19	.08	.09	.03	.00	.06	.04	.00
3	3	1.237	1.248	.03	.00	.00	.24	.26	.04	.17	.00
4	4	.987	1.396	.12	.02	.01	.17	.03	.51	.03	.00
5	5	.874	1.484	.03	.00	.01	.27	.65	.01	.10	.00
6	6	.729	1.625	.63	.11	.19	.00	.05	.06	.01	.00
7	7	.600	1.792	.00	.00	.09	.27	.00	.29	.59	.08
8	8	.223	2.939	.00	.73	.55	.02	.01	.01	.02	.82

a. Dependent Variable: y

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-61

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΜΕΤΑΒΑΛΟΥΜΕ ΤΩΡΑ ΤΗΝ ΤΕΛΕΙΑ ΓΡΑΜΜΙΚΗ ΣΧΕΣΗ ΣΕ ΠΟΛΥ ΥΨΗΛΗ

ΔΗΛΩΣΗ

COMPUTE e2 = RV.NORMAL(0,0.1) .
EXECUTE .

COMPUTE x1 = x2+x3+x4 + e2 .
EXECUTE .

ΞΑΝΑΚΑΝΟΥΜΕ ΤΗΝ ΠΑΛΙΝΔΡΟΜΗΣΗ ΤΩΝ Χ ΣΤΗΝ Υ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-62

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta				Tolerance	VIF
1	(Constant)	4.031	.044			90.649	.000		
	x1	-.197	.449	-.055	-.440	.661	.004	.277	.006
	x2	1.162	.450	.206	2.582	.011	.009	.115	.060
	x3	.219	.451	.040	.486	.628	.008	.120	.070
	x4	-2.814	.454	-.449	-6.203	.000	.011	.94	.506
	x5	-.026	.052	-.004	-.503	.616	.924	1.082	
	x6	4.925	.044	.866	111.230	.000	.916	1.092	
	x7	.047	.051	.007	.921	.360	.906	1.104	
	x8	-.024	.048	-.004	-.509	.612	.844	1.185	

a. Dependent Variable: y

ΠΡΑΓΜΑΤΙΚΗ ΣΧΕΣΗ: $Y = 4 + X_2 - 3 X_4 + 5 X_6 + \epsilon$, $\epsilon \sim N(0, 0.25 = 0.5^2)$
ΕΚΤΙΜΩΜΕΝΗ ΣΧΕΣΗ:
 $Y = 4 - 0.2 X_1 + 1.2 X_2 + 0.22 X_3 - 2.8 X_4 - 0.02 X_5 + 4.9 X_6 + 0.05 X_7 - 0.02 X_8 + \epsilon$,
 $\epsilon \sim N(0, (0.42)^2)$

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-63

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	x1	x2	x3	x4	x5	x6	x7
1	1	1.998	1.000	.00	.00	.00	.00	.00	.00	.02	.03
2	2	1.484	1.160	.16	.00	.00	.00	.00	.05	.00	.04
3	3	1.271	1.254	.00	.00	.00	.00	.00	.14	.22	.10
4	4	1.101	1.347	.06	.00	.00	.00	.00	.11	.02	.06
5	5	.985	1.425	.16	.00	.00	.00	.00	.13	.04	.06
6	6	.871	1.514	.05	.00	.00	.00	.00	.28	.61	.01
7	7	.720	1.666	.56	.00	.00	.00	.00	.01	.06	.03
8	8	.571	1.870	.00	.00	.00	.00	.00	.27	.00	.31
9	9	.002	34.993	.00	1.00	.99	.99	.99	.00	.05	.00

a. Dependent Variable: y

ΣΤΗΝ ΠΡΟΒΛΗΜΑΤΙΚΗ ΙΔΙΟΤΗΤΗ ΑΝΤΙΣΤΟΙΧΕΙ ΓΡΑΜΜΙΚΟΣ ΣΥΝΔΙΑΣΜΟΣ ΣΤΟΝ ΟΠΟΙΟ ΣΥΜΜΕΤΕΧΟΥΝ ΜΕ ΜΕΓΑΛΗ ΠΟΣΟΣΤΑ (ΣΧΕΔΟΝ 100%) ΟΙ X_1, X_2, X_3 ΚΑΙ X_4 .

1 ΜΙΚΡΗ (ΠΡΟΒΛΗΜΑΤΙΚΗ) ΙΔΙΟΤΗΤΗ ΜΕ $CP > 30 \Leftrightarrow$ ΣΥΝΕΠΩΣ ΥΠΑΡΧΕΙ ΕΝΑΣ ΓΡΑΜΜΙΚΟΣ ΣΥΝΔΙΑΣΜΟΣ (ΣΧΕΣΗ) ΜΕΤΑΞΥ ΤΩΝ $X \Leftrightarrow$ ΑΡΑ ΠΡΕΠΕΙ ΝΑ ΑΦΑΙΡΕΣΟΥΜΕ 1 ΜΕΤΑΒΛΗΤΗ ΜΟΝΟ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-64

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΑΝ ΒΑΛΟΥΜΕ BACKWARD SELECTION

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	x8, x3, x6, x7, x5, x2, x4, x1		Enter
2		x1	Backward (criterion: Probability of F-to-remove >= .100).
3		x8	Backward (criterion: Probability of F-to-remove >= .100).
4		x3	Backward (criterion: Probability of F-to-remove >= .100).
5		x5	Backward (criterion: Probability of F-to-remove >= .100).
6		x7	Backward (criterion: Probability of F-to-remove >= .100).

- a. All requested variables entered.
- b. Dependent Variable: y

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΑΝ ΒΑΛΟΥΜΕ BACKWARD SELECTION

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997 ^a	.995	.995	4.2489
2	.997 ^b	.995	.995	4.2302
3	.997 ^c	.995	.995	4.2134
4	.997 ^d	.995	.995	4.1978
5	.997 ^e	.995	.995	4.1851
6	.997 ^f	.995	.995	4.1808

ΜΕΤΑΒΟΛΗ R, R² & σ ΑΝΑ ΒΗΜΑ
(ΔΗΛΑΔΗ ΓΙΑ ΚΑΘΕ ΜΟΝΤΕΛΟ ΜΕΤΑ
ΑΠΟ ΤΗΝ ΣΤΑΔΙΑΚΗ ΑΦΑΙΡΕΣΗ
ΜΕΤΑΒΛΗΤΩΝ)

- a. Predictors: (Constant), x8, x3, x6, x7, x5, x2, x4, x1
- b. Predictors: (Constant), x8, x3, x6, x7, x5, x2, x4
- c. Predictors: (Constant), x3, x6, x7, x5, x2, x4
- d. Predictors: (Constant), x6, x7, x5, x2, x4
- e. Predictors: (Constant), x6, x7, x2, x4
- f. Predictors: (Constant), x6, x2, x4

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΑΝ ΒΑΛΟΥΜΕ BACKWARD SELECTION

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		Collinearity Statistics	
		B	Std. Error	Beta		Tolerance	VIF
5	(Constant)	4.030	.044				
	x2	-.967	.042	.172	22.773	.000	.947
	x4	-3.013	.046	-.481	-65.301	.000	.992
	x6	4.936	.042	.868	117.703	.000	.990
	x7	.044	.049	.007	.895	.373	.947
	6	(Constant)	4.030	.044			
6	x2	.975	.041	.173	23.516	.000	.991
	x4	-3.010	.046	-.480	-65.516	.000	.999
	x6	4.934	.042	.868	117.877	.000	.992

a. Dependent Variable: y

ΠΡΑΓΜΑΤΙΚΗ ΣΧΕΣΗ: $Y = 4 + 1.00 X_2 - 3 X_4 + 5.0 X_6 + \epsilon$, $\epsilon \sim N(0, (0.50)^2)$
ΕΚΤΙΜΩΜΕΝΗ ΣΧΕΣΗ: $Y = 4 + 0.98 X_2 - 3 X_4 + 4.9 X_6 + \epsilon$, $\epsilon \sim N(0, (0.42)^2)$

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΑΝ ΒΑΛΟΥΜΕ BACKWARD SELECTION

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions													
				(Constant)	x1	x2	x3	x4	x5	x6	x7	x8					
6	1	1.319	1.000														
	2	1.960	1.115	.02	.22	.22	.05	.05	.03	.03	.03	.03	.03	.03	.03	.03	.03
	3	.888	1.219	.00	.48	.48	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28	.28
	4	.733	1.341	.67	.15	.15	.15	.15	.15	.15	.15	.15	.15	.15	.15	.15	.15

OK

Excluded Variables^b

ΕΛΑΧΙΣΤΟ TOLERANCE ΑΝ ΠΡΟΣΤΕΘΕΙ
ΑΥΤΗ Η ΜΕΤΑΒΛΗΤΗ ΣΤΟ ΜΟΝΤΕΛΟ

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
6	x1	-.006 ^a	-.547	.586	.056	.419	2.387	.419
	x8	-.004 ^a	-.490	.625	-.050	.898	1.114	.898
	x3	-.004 ^a	-.591	.556	.061	.967	1.034	.967
	x5	-.005 ^a	-.720	.473	-.074	.961	1.040	.961
	x7	.007 ^a	.895	.373	.091	.947	1.056	.947

Collinearity diagnostics της κάθε μεταβλητής αν προστεθεί στο μοντέλο

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΑΝ ΒΑΛΟΥΜΕ STEPWISE PROCEDURE

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	x6		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	x4		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	x2		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	.854 ^a	.729	.726	2.99901
2	.982 ^b	.965	.964	1.08144
3	.997 ^c	.995	.995	.41808

a. Predictors: (Constant), x6
 b. Predictors: (Constant), x6, x4
 c. Predictors: (Constant), x6, x4, x2

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-69

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΑΝ ΒΑΛΟΥΜΕ STEPWISE PROCEDURE

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.193	.300		10.641	.000		
	x6	4.856	.299	.854	16.238	.000	1.000	1.000
2	(Constant)	3.882	.112		34.815	.000		
	x6	4.846	.108	.852	44.936	.000	1.000	1.000
	x4	-3.044	.119	-.486	-25.626	.000	1.000	1.000
3	(Constant)	4.030	.044		92.513	.000		
	x6	4.934	.042	.868	117.877	.000	.992	1.008
	x4	-3.010	.046	-.480	-65.516	.000	.999	1.001
	x2	-.975	.041	-.173	-23.516	.000	.991	1.009

a. Dependent Variable: y

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-70

Excluded Variables^a

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics			
						Tolerance	VIF	Minimum Tolerance	
1	x1	-.109 ^a	-2.112	.037	-.210	.998	1.002	.998	
	x2	.188 ^a	3.803	.000	.360	.992	1.008	.992	
	x3	.082 ^a	1.570	.120	.157	.998	1.002	.998	
	x4	-.486 ^a	-25.626	.000	-.933	1.000	1.000	1.000	
	x5	-.072 ^a	-1.358	.178	-.137	.985	1.015	.985	
	x7	.007 ^a	.129	.898	.013	.997	1.003	.997	
	x8	.103 ^a	1.983	.050	.197	.991	1.009	.991	
	2	x1	.136 ^b	8.414	.000	.651	.797	1.254	.797
x2		-.173 ^b	-23.516	.000	-.923	.991	1.009	.991	
x3		-.002 ^b	-.111	.912	-.011	.968	1.033	.968	
x5		.001 ^b	.071	.944	.007	.963	1.039	.963	
x7		.043 ^b	2.309	.023	.229	.991	1.009	.991	
x8		-.031 ^b	-1.578	.118	-.159	.919	1.088	.919	
3		x1	.006 ^c	.547	.586	.056	.419	2.387	.419
		x3	.004 ^c	.591	.556	.061	.967	1.034	.967
	x5	-.005 ^c	-.720	.473	-.074	.961	1.040	.961	
	x7	.007 ^c	.895	.373	.091	.947	1.056	.947	
x8	-.004 ^c	-.490	.625	-.050	.898	1.114	.898		

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-71

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-2 – ΣΥΓΓΡΑΜΜΙΚΑ ΔΕΔΟΜΕΝΑ)

ΜΟΝΤΕΛΟ: $Y = \beta_0 + \beta_6 X_6$

ΠΑΛΙΝΔΡΟΜΗΣΗ X_4 & X_6 ΣΤΗΝ Y

$Y = \beta_0 + \beta_6 X_6 + \beta_1 X_1$

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.209	.295		10.876	.000		
	x6	4.828	.294	.869	16.410	.000	.998	1.002
2	(Constant)	3.96	.187		21.166	.000		
	x1	-.109	.037	-.210	-2.112	.037	.998	1.002

a. Dependent Variable: y

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II -72

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

Ποια στοιχεία - δείκτες καθορίζουν το το γενικό οικονομικό
επίπεδο μιας χώρας;

ΒΗΜΑΤΑ

1. ΕΠΙΛΟΓΗ Y – HISTOGRAM
2. ΕΠΙΛΟΓΗ X – BACKWARD/STEPWISE SELECTION
3. ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ
4. ΕΡΜΗΝΕΙΑ ΤΕΛΙΚΟΥ ΜΟΝΤΕΛΟΥ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνει 11-73

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

1... ΕΠΙΛΟΓΗ Y
GDP_CAP = Gross domestic product / capita
Ακαθάριστο εθνικό προϊόν ανά άτομο
Y = log(GDP) λόγω ασυμμετρίας

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνει 11-74

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

2... ΕΠΙΛΟΓΗ X – ΑΡΧΙΚΗ ΕΠΙΛΟΓΗ X

■ populatn	11.birth_rt
■ density	12.death_rt
■ urban	13.aids_rt
■ lifeexpf	14.lg_aidsr
■ lifeexpm	15.b_to_d
■ literacy	16.fertilty
■ pop_incr	17.log_pop
■ babymort	18.croprgrow
■ calories	19.lit_male
■ aids	20.lit_fema

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνει 11-75

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

2... ΕΠΙΛΟΓΗ X – BACKWARD SELECTION

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνει 11-76

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

2... ΕΠΙΛΟΓΗ Χ – BACKWARD SELECTION

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
2	lg_aidsr	Log (base 10) of AIDS_RT	Backward (criterion: Probability of F-to-remove <= .050)
3	cropgrow	Population increase (% per year)	Backward (criterion: Probability of F-to-remove <= .050)
4	sop_incr	Death rate per 1000 people	Backward (criterion: Probability of F-to-remove <= .050)
5	death_rt	People who read (%)	Backward (criterion: Probability of F-to-remove <= .050)
6	literacy	Number of people / sq. kilometer	Backward (criterion: Probability of F-to-remove <= .050)
7	density	Birth to death ratio	Backward (criterion: Probability of F-to-remove <= .050)
8	b_to_d	Average male life expectancy	Backward (criterion: Probability of F-to-remove <= .050)
9	lifeexpm	Infant mortality (deaths per 1000 live births)	Backward (criterion: Probability of F-to-remove <= .050)
10	babymort	Population in thousands	Backward (criterion: Probability of F-to-remove <= .050)
11	populatn	Males who read (%)	Backward (criterion: Probability of F-to-remove <= .050)
12	lit_male	Aids cases	Backward (criterion: Probability of F-to-remove <= .050)
13	aids	Females who read (%)	Backward (criterion: Probability of F-to-remove <= .050)
14	lit_fema	Average female life expectancy	Backward (criterion: Probability of F-to-remove <= .050)
15	lifeexpf	Number of aids cases / 100000 people	Backward (criterion: Probability of F-to-remove <= .050)
16	aids_rt		

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-77

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

2... ΕΠΙΛΟΓΗ Χ – STEPWISE SELECTION

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	calories	Daily calorie intake	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051)
2	urban	People living in cities (%)	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051)
3	log_pop	Log (base 10) of Population	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051)
4	birth_rt	Birth rate per 1000 people	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051)
5	aids	Aids cases	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051)
6	fertility	Fertility: average number of kids	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051)

^a. Dependent Variable: log_gdp Log (base 10) of GDP_CAP

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-78

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

**BACKWARD ME P-TO-REMOVE = 0.05
ΜΕΤΑΒΛΗΤΕΣ ΣΤΟ ΜΟΝΤΕΛΟ**

- 1) urban People living in cities (%)
- 2) calories Daily calorie intake
- 3) fertility Fertility: average number of kids
- 4) birth_rt Birth rate per 1000 people
- 5) log_pop Log (base 10) of Population

**STEPWISE ME P-TO-REMOVE = 0.051/ P-TO-ADD 0.050
ΜΕΤΑΒΛΗΤΕΣ ΣΤΟ ΜΟΝΤΕΛΟ**

- ΟΙ ΠΑΡΑΠΑΝΩ + AIDS

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-79

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική 11.5. Το πρόβλημα της πολυσυγγραμμικότητας (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

BACKWARD

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.936 ^a	.875	.814	24528
2	.936 ^b	.875	.814	24213
3	.936 ^b	.875	.819	23913
4	.936 ^d	.875	.823	23624
5	.935 ^d	.875	.827	23347
6	.935 ^d	.875	.831	23087
7	.935 ^d	.874	.835	22862
8	.935 ^d	.873	.837	22706
9	.934 ^f	.872	.838	22611
10	.931 ^f	.867	.836	22761
11	.928 ^h	.862	.833	22987
12	.925 ^h	.856	.830	23204
13	.922 ^m	.849	.825	23495
14	.916 ⁿ	.840	.818	23999
15	.911 ⁿ	.830	.810	24487
16	.908 ^p	.824	.808	24647

STEPWISE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.803 ^a	.644	.638	33805
2	.869 ^b	.754	.746	28344
3	.883 ^c	.779	.767	27126
4	.899 ^d	.808	.794	25510
5	.911 ^e	.829	.813	24312
6	.916 ^f	.842	.824	23586

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-80

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

EXCLUDED VARIABLES*						
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
16	lg_aidsr Log (base 10) of AIDS_RT	.088 ^a	1.411	.164	.192	.895
	popgrw	-.081 ^a	-1.148	.256	-.157	.666
	pop_inc Population increase (% per year)	.012 ^a	.117	.907	.016	.307
	death_rt Death rate per 1000 people	-.029 ^a	-.494	.623	-.069	.547
	literacy People who read (%)	.056 ^a	.515	.609	.071	.291
	density Number of people / sq kilometer	.000 ^a	-.001	.999	.000	.944
	b_to_d Birth to death ratio	-.008 ^a	-.116	.908	-.016	.727
	lifeexpm Average male life expectancy	.072 ^a	.641	.524	.089	.269
	babymort Infant mortality (deaths per 1000 live births)	-.078 ^a	-.647	.520	-.089	.233
	populate Population in thousands	-.061 ^a	-.787	.447	-.108	.523
	lit_male Males who read (%)	.001 ^a	.013	.990	.003	.366
	aidsr Aids cases	.144 ^a	2.424	.019	.319	.865
	lit_femr Females who read (%)	-.010 ^a	-.089	.929	-.012	.272
	lifeexpf Average female life expectancy	.113 ^a	.890	.377	.123	.205
	aids_f Number of aids cases / 100000 people	.080 ^a	1.302	.199	.178	.873

ΣΤΑΤΙΣΤΙΚΑ ΤΩΝ EXCLUDED VARIABLES ΑΠΟ ΤΗ BACKWARD PROCEDURE

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνειο 11-81

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΣΥΝΕΠΩΣ ΔΟΥΛΕΥΟΥΜΕ ΤΩΡΑ ΜΕ ΤΟ ΤΕΛΙΚΟ ΜΟΝΤΕΛΟ ΠΟΥ ΕΠΙΛΕΧΘΗΚΕ ΑΠΟ ΤΗ STEPWISE PROCEDURE

urban
calories
fertility
birth_rt
log_pop
AIDS

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνειο 11-83

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

EXCLUDED VARIABLES*						
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
16	lg_aidsr Log (base 10) of AIDS_RT	.088 ^a	1.411	.164	.192	.895
	popgrw	-.081 ^a	-1.148	.256	-.157	.666
	pop_inc Population increase (% per year)	.012 ^a	.117	.907	.016	.307
	death_rt Death rate per 1000 people	-.029 ^a	-.494	.623	-.069	.547
	literacy People who read (%)	.056 ^a	.515	.609	.071	.291
	density Number of people / sq kilometer	.000 ^a	-.001	.999	.000	.944
	b_to_d Birth to death ratio	-.008 ^a	-.116	.908	-.016	.727
	lifeexpm Average male life expectancy	.072 ^a	.641	.524	.089	.269
	babymort Infant mortality (deaths per 1000 live births)	-.078 ^a	-.647	.520	-.089	.233
	populate Population in thousands	-.061 ^a	-.787	.447	-.108	.523
	lit_male Males who read (%)	.001 ^a	.013	.990	.003	.366
	aidsr Aids cases	.144 ^a	2.424	.019	.319	.865
	lit_femr Females who read (%)	-.010 ^a	-.089	.929	-.012	.272
	lifeexpf Average female life expectancy	.113 ^a	.890	.377	.123	.205
	aids_f Number of aids cases / 100000 people	.080 ^a	1.302	.199	.178	.873

ΣΤΑΤΙΣΤΙΚΑ ΤΩΝ EXCLUDED VARIABLES ΑΠΟ ΤΗ BACKWARD PROCEDURE

ΑΡΑ ΚΡΑΤΑΜΕ ΤΟ ΜΟΝΤΕΛΟ ΤΗΣ STEPWISE ΔΙΑΔΙΚΑΣΙΑΣ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνειο 11-82

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
11.5. Το πρόβλημα της πολυσυγγραμμικότητας
(ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΣΥΝΕΠΩΣ ΔΟΥΛΕΥΟΥΜΕ ΤΩΡΑ ΜΕ ΤΟ ΤΕΛΙΚΟ ΜΟΝΤΕΛΟ ΠΟΥ ΕΠΙΛΕΧΘΗΚΕ ΑΠΟ ΤΗ STEPWISE PROCEDURE

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.258	.397		8.203	.000		
	urban People living in cities (%)	.008	.002	.298	4.748	.000	.453	2.209
	calories Daily calorie intake	.000	.000	.323	4.275	.000	.311	3.212
	log_pop Log (base 10) of Population	-.179	.050	-.169	-3.622	.001	.820	1.220
	aids Aids cases	1.18E-006	.000	.086	1.908	.061	.886	1.128
	fertility Fertility: average number of kids	.221	.068	.645	3.265	.002	.046	21.883
	birth_rt Birth rate per 1000 people	-.054	.012	-.1002	-4.643	.000	.038	26.157

^a. Dependent Variable: log_gdp Log (base 10) of GDP_CAP

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνειο 11-84

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΣΥΝΕΠΩΣ ΔΟΥΛΕΥΟΥΜΕ ΤΩΡΑ ΜΕ ΤΟ ΤΕΛΙΚΟ ΜΟΝΤΕΛΟ ΠΟΥ ΕΠΙΛΕΧΘΗΚΕ ΑΠΟ ΤΗ STEPWISE PROCEDURE

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	urban People living in cities (%)	calories Daily calorie intake	log_pop Log (base 10) of Population	aids Aids cases	fertility Fertility: average number of kids	birth_rt Birth rate per 1000 people	
1	1	5.511	1.000	.00	.00	.00	.00	.00	.00	.00	.00
2	2	.949	2.410	.00	.00	.00	.00	.85	.00	.00	.00
3	3	.453	3.487	.00	.08	.00	.00	.04	.01	.01	.00
4	4	.062	9.400	.00	.56	.01	.09	.02	.01	.00	.00
5	5	.016	18.665	.00	.35	.43	.35	.00	.02	.01	.00
6	6	.006	31.581	.33	.04	.04	.48	.06	.44	.23	.00
7	7	.003	44.258	.67	.00	.52	.07	.02	.52	.75	.00

^a Dependent Variable: log_gdp Log (base 10) of GDP_CAP

Αφαιρούμε το Birth διότι έχει μεγαλύτερο VIF+μεγαλύτερη συμμετοχή στον γρ. Συνδιασμό με τη μικρότερη ιδιοτιμή

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-85

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ BIRTH_RT

R²: 0.88 -> 0.84

R_{adj}²: 0.87->0.83

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error				Beta	Tolerance
1	(Constant)	2.446	.407		6.011	.000		
	urban People living in cities (%)	.008	.002	.295	4.121	.000	.453	2.209
	calories Daily calorie intake	.001	.000	.484	6.299	.000	.393	2.542
	log_pop Log (base 10) of Population	-.195	.056	-.183	-3.456	.001	.823	1.215
	aids Aids cases	1.19E-006	.000	.087	1.697	.094	.886	1.128
	fertility Fertility: average number of kids	-.077	.025	-.225	-3.137	.003	.453	2.209

^a Dependent Variable: log_gdp Log (base 10) of GDP_CAP

OK

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-86

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ BIRTH_RT

R²: 0.88 -> 0.84

R_{adj}²: 0.87->0.83

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	urban People living in cities (%)	calories Daily calorie intake	log_pop Log (base 10) of Population	aids Aids cases	fertility Fertility: average number of kids		
1	1	4.665	1.000	.00	.00	.00	.00	.00	.00	.00	.00
2	2	.932	2.237	.00	.00	.00	.00	.88	.00	.00	.00
3	3	.325	3.800	.00	.08	.00	.00	.01	.17	.00	.00
4	4	.059	8.858	.01	.52	.01	.09	.02	.31	.00	.00
5	5	.015	17.454	.00	.38	.59	.40	.00	.01	.00	.00
6	6	.004	32.210	.99	.02	.40	.51	.08	.50	.00	.00

^a Dependent Variable: log_gdp Log (base 10) of GDP_CAP

Υπάρχουν CP μεγάλα όμως στο χειρότερο γρ. Συνδιασμό δεν εμπλέκεται με μεγάλο ποσοστό κάποια από τις Χ

Και επειδή δεν έχουμε μεγάλα VIF προχωράμε με αυτό το μοντέλο

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-87

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ BIRTH_RT

Tests of Normality

ZRE_2 Standardized Residual	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
	.128	74	.004	.964	74	.035

^a Lilliefors Significance Correction

ΒΙΟΣΤΑΤΙΣΤΙΚΗ II Διαφάνεια 11-88

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ FERTILITY (OXI TO BIRTH_RATE)
R²: 0.88 -> 0.84 -> 0.86
R_{adj}²: 0.87-> 0.83 -> 0.85

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.989	.415		7.200	.000		
	urban People living in cities (%)	.007	.002	.276	4.144	.000	.458	2.185
	calories Daily calorie intake	.000	.000	.399	5.180	.000	.343	2.913
	aids Aids cases	1.26E-006	.000	.092	1.915	.060	.888	1.126
	log_pop Log (base 10) of Population	-.204	.052	-.192	-3.891	.000	.838	1.193
	birth_rt Birth rate per 1000 people	-.018	.004	-.334	-4.557	.000	.379	2.640

^a. Dependent Variable: log_gdp Log (base 10) of GDP_CAP
OK

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-89

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ FERTILITY (OXI TO BIRTH_RATE)
R²: 0.88 -> 0.84 -> 0.86
R_{adj}²: 0.87-> 0.83 -> 0.85

Collinearity Diagnostics ^a									
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	urban People living in cities (%)	calories Daily calorie intake	aids Aids cases	log_pop Log (base 10) of Population	birth_rt Birth rate per 1000 people
1	1	4.702	1.000	.00	.00	.00	.00	.00	.00
	2	.933	2.245	.00	.00	.00	.88	.00	.00
	3	.294	4.002	.00	.10	.00	.01	.00	.12
	4	.053	9.429	.00	.55	.02	.02	.10	.27
	5	.015	17.786	.00	.33	.47	.01	.49	.02
	6	.004	35.101	.99	.01	.51	.07	.41	.59

^a. Dependent Variable: log_gdp Log (base 10) of GDP_CAP
 Υπάρχουν CP μεγάλα όμως στο χειρότερο γρ. Συνδιασμό δεν εμπέκεται με μεγάλο ποσοστό κάποια από τις X
 Και επειδή δεν έχουμε μεγάλο VIF προχωράμε με αυτό το μοντέλο

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-90

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ FERTILITY (OXI TO BIRTH_RATE)

Tests of Normality					
ZRE_S Standardized Residual	Kolmogorov-Smirnov ^a		Shapiro-Wilk		
	Statistic	df	Statistic	df	Sig.
	.083	74	.200*	.973	.74

^a. This is a lower bound of the true significance.
^a. Lilliefors Significance Correction
 OK ΚΑΝΟΝΙΚΟΤΗΤΑ

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-91

11. Επίδραση πολλών μεταβλητών σε μια ποσοτική
 11.5. Το πρόβλημα της πολυσυγγραμμικότητας
 (ΠΑΡΑΔΕΙΓΜΑ 11-3 – WORLD 95)

ΑΦΑΙΡΟΥΜΕ ΤΟ FERTILITY (OXI TO BIRTH_RATE)

ΠΕΡΙΜΕΝΟΥΜΕ ΤΟ 5% ΤΩΝ ΤΙΜΩΝ ΝΑ ΕΙΝΑΙ ΕΚΤΟΣ ΟΡΙΩΝ ΔΗΛ. 0.05*74 = 4 (ΠΕΡΙΠΟΥ)
 ⇨ OK

ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΙΙ Διαφάνεια 11-92

