ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

6. Συσχέτιση

Γενικά

- Υπάρχει σχέση ανάμεσα σε δύο (ή περισσότερες) μεταβλητές;
- Αν υπάρχει σχέση ποια η φύση της σχέσης αυτής;
- Συσχέτιση: μέτρο σχέσης ανάμεσα σε μεταβλητές
  - Θετικά συσχετισμένες
  - Αρνητικά συσχετισμένες
  - Ασυσχέτιστες

Μέτρηση μεταβλητότητας μιας μεταβλητής – διασπορά

- **Διασπορά** ή **διακύμανση** (*variance*) μιας μεταβλητής:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

- Ερμηνεία: το μέσο ποσό μεταβλητότητας των παρατηρήσεων x_i από τη μέση τιμή $\bar{x}$

Μέτρηση συμμεταβλητότητας - συνδιασπορά

- **Συνδιασπορά** ή **συνδιακύμανση** (*covariance*) δύο μεταβλητών

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Ερμηνεία: Το μέσο ποσό της «ταυτόχρονης» μεταβλητότητας των x και y από τις μέσες τιμές τους

Η συνδιασπορά ως μέτρο σχέσης

- Κεντρική ιδέα:
  - *Αν πράγματι οι δύο μεταβλητές σχετίζονται, τότε όπως μεταβάλλεται η μια (x) γύρω από τη μέση τιμή της, με παρόμοιο τρόπο (ή με ακριβώς αντίθετο τρόπο) θα μεταβάλλεται και η άλλη (y) γύρω από τη μέση τιμή της.*

Παράδειγμα

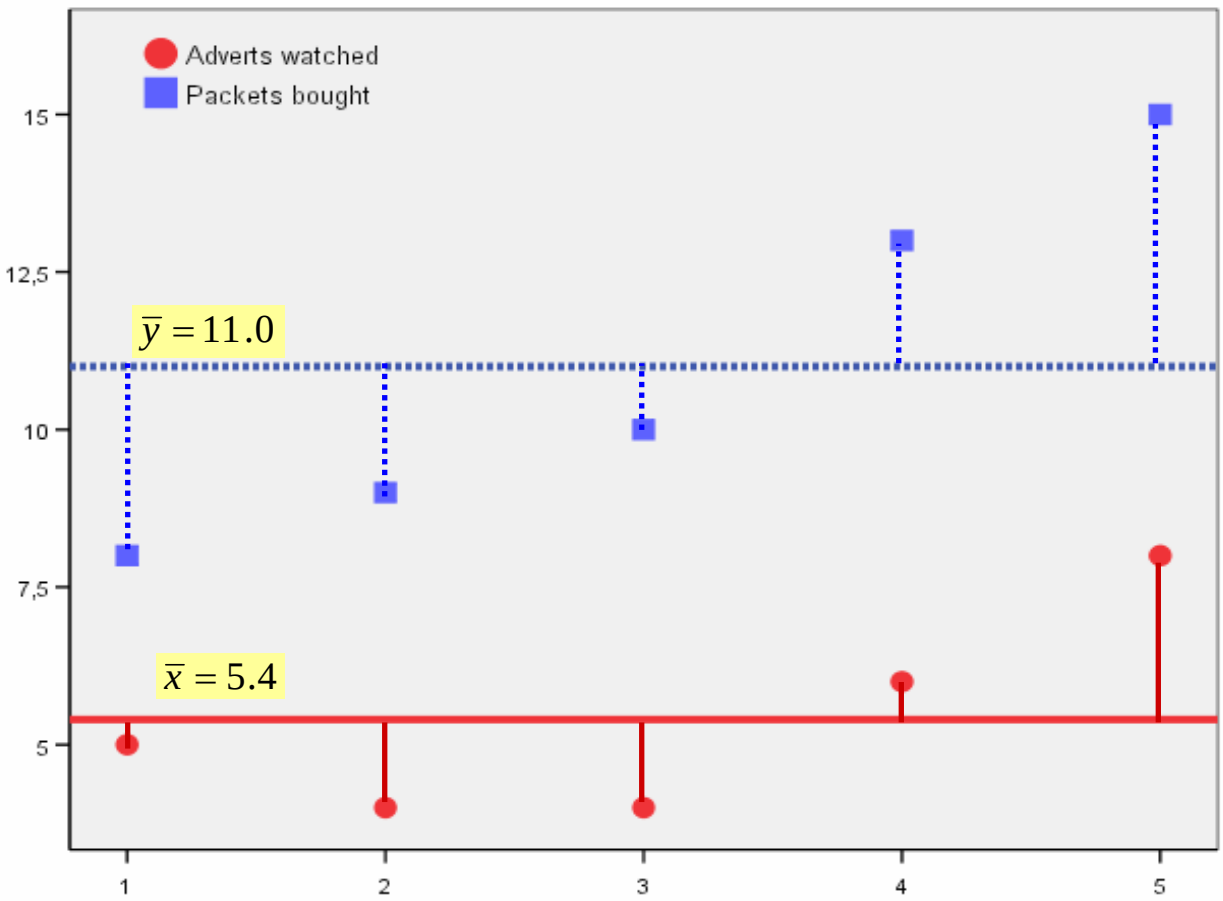
- Σε δείγμα 5 ατόμων προβλήθηκε ένας αριθμός από διαφημίσεις για μια σοκολάτα και την επόμενη εβδομάδα μετρήθηκε πόσες σοκολάτες αγόρασαν

Case Summaries[a]

		Adverts watched	Packets bought
1		5	8
2		4	9
3		4	10
4		6	13
5		8	15
Total	N	5	5
	Mean	5,40	11,00
	Std. Deviation	1,673	2,915

a. Limited to first 100 cases.

Διαφορά τιμών των μεταβλητών από τις μέσες τιμές τους



Γενικά παρατηρούμε ομοιότητα στη συμπεριφορά των μεταβλητών ως προς τη μεταβλητότητά τους γύρω από τις μέσες τιμές τους

Υπολογισμός της συνδιασποράς

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} =$$

$$\frac{(-.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (-.6)(2) + (2.6)(4)}{4} = 4.25$$

- **Θετική συνδιασπορά:** οι μεταβλητές μεταβάλλονται προς την ίδια κατεύθυνση από τη μέση τιμή τους
- **Αρνητική συνδιασπορά:** οι μεταβλητές μεταβάλλονται προς την αντίθετη κατεύθυνση από τη μέση τιμή τους
- Πρόβλημα: Πως καταλαβαίνουμε αν η συνδιασπορά (και επομένως η σχέση) είναι μεγάλη;

Συντελεστής συσχέτισης

- **Τυποποίηση** (*standardization*) της συνδιασποράς
- Απαλλαγή του μέτρου από μονάδες μέτρησης – διαίρεση με τυπικές αποκλίσεις των μεταβλητών
- **Συντελεστής συσχέτισης του Pearson** (*Pearson correlation coefficient*):

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$$

Τιμές και ερμηνεία του συντελεστή συσχέτισης

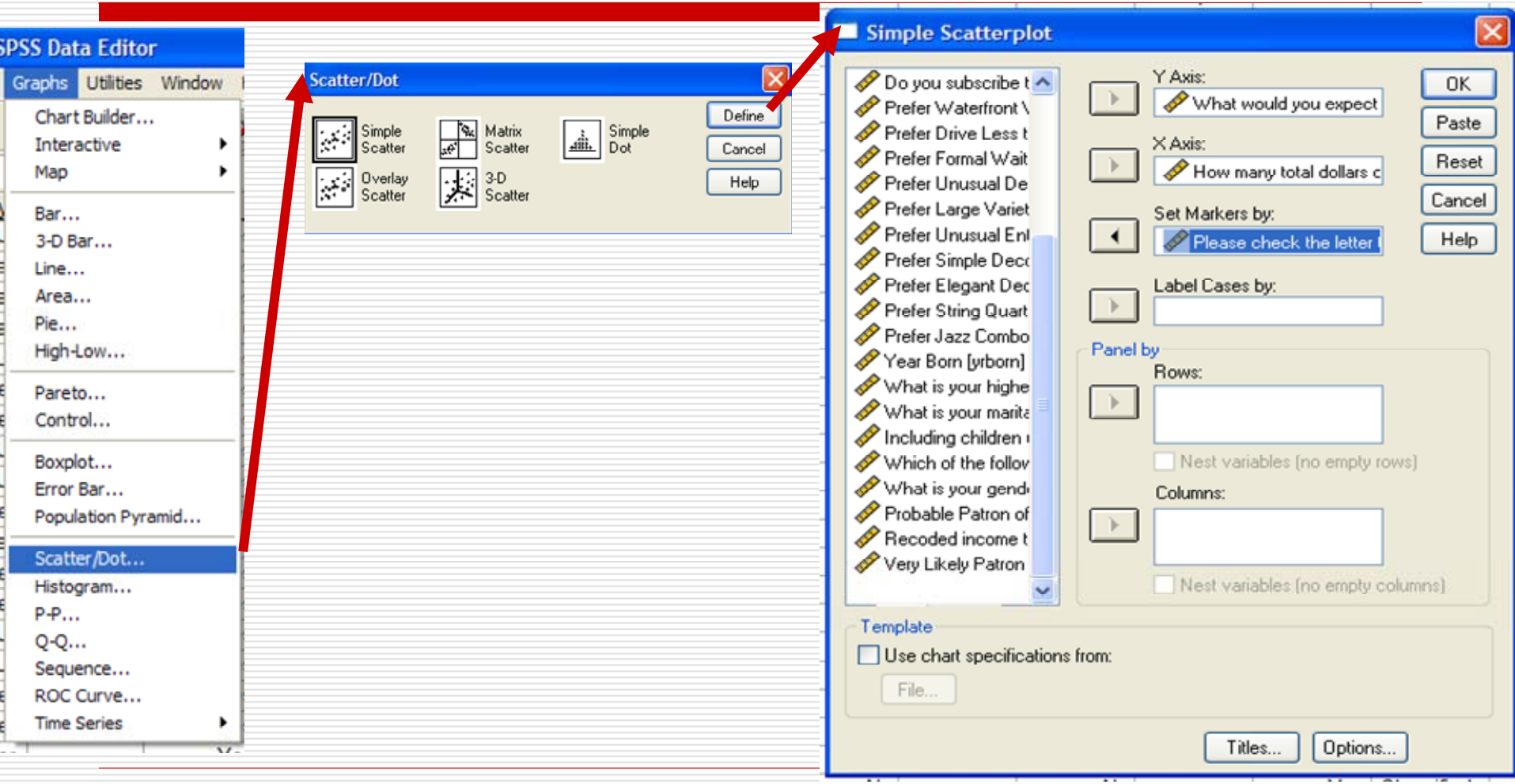
- Οι τιμές του r είναι πάντοτε στο διάστημα [-1, +1]
- r=+1: Οι μεταβλητές είναι **θετικά συσχετισμένες** (όταν η μια αυξάνει, η άλλη αυξάνει γραμμικά)
- r=-1: Οι μεταβλητές είναι **αρνητικά συσχετισμένες** (όταν η μια αυξάνει, η άλλη μειώνεται γραμμικά)
- r=0: Οι μεταβλητές είναι **ασυσχέτιστες**

Τιμές και ερμηνεία του συντελεστή συσχέτισης (συν.)

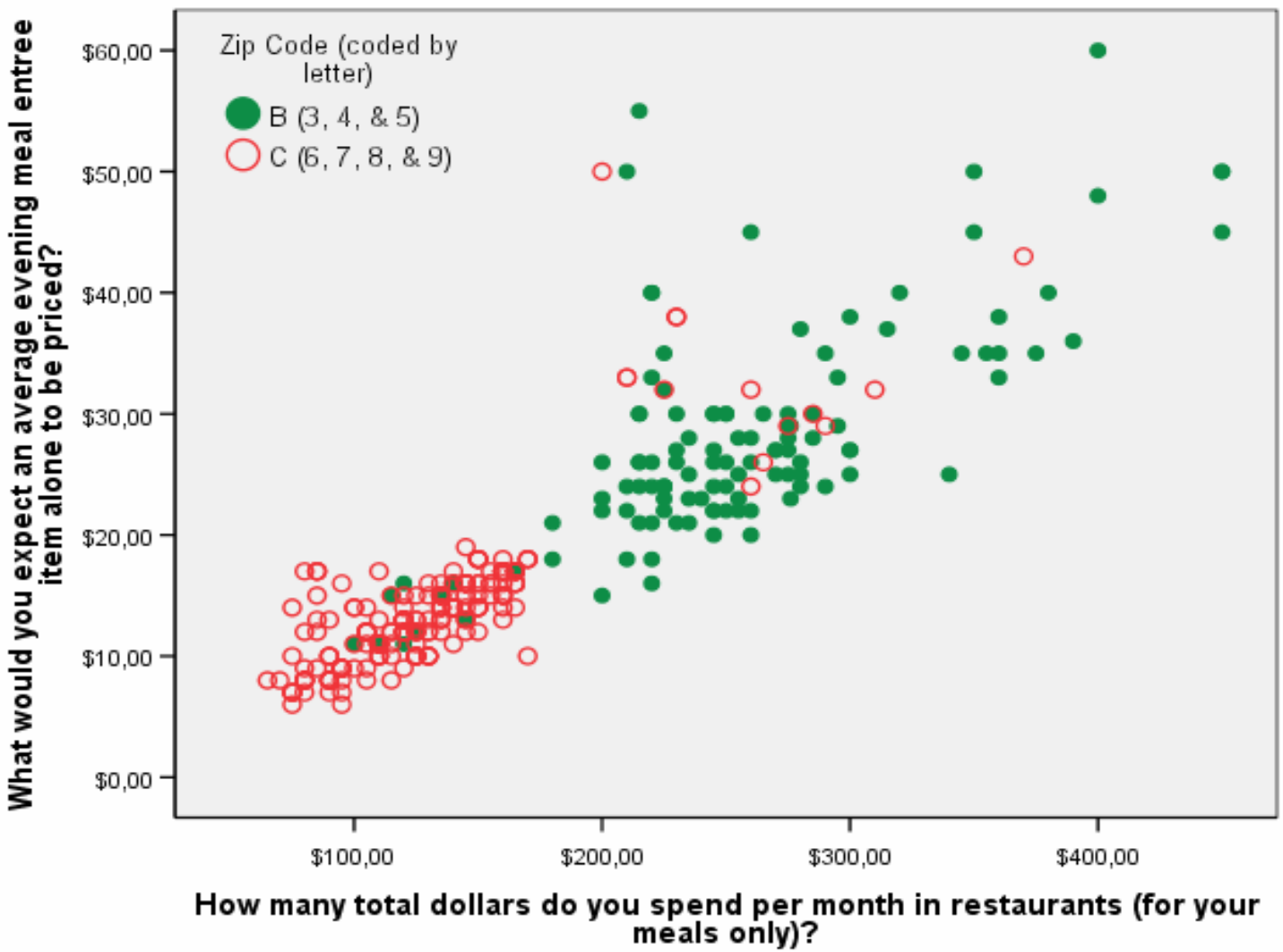
- Εμπειρικός κανόνας:
  - r=±0.1: μικρή συσχέτιση
  - r=±0.3: μέτρια συσχέτιση
  - r=±0.5: ισχυρή συσχέτιση
- Στο παράδειγμα ισχυρή συσχέτιση:

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} = \frac{4.25}{(1.67)(2.92)} = 0.87$$

Γραφική παράσταση της συσχέτισης: Το διάγραμμα διασποράς



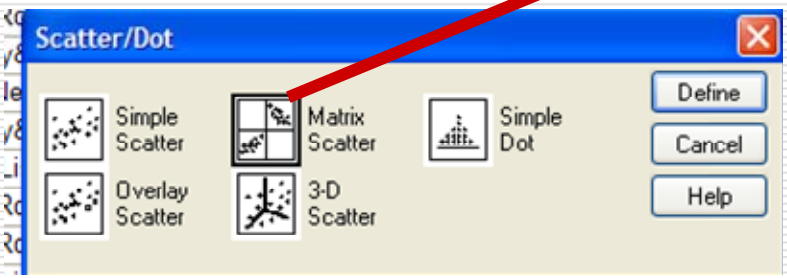
Απλό διάγραμμα διασποράς (simple scatterplot)

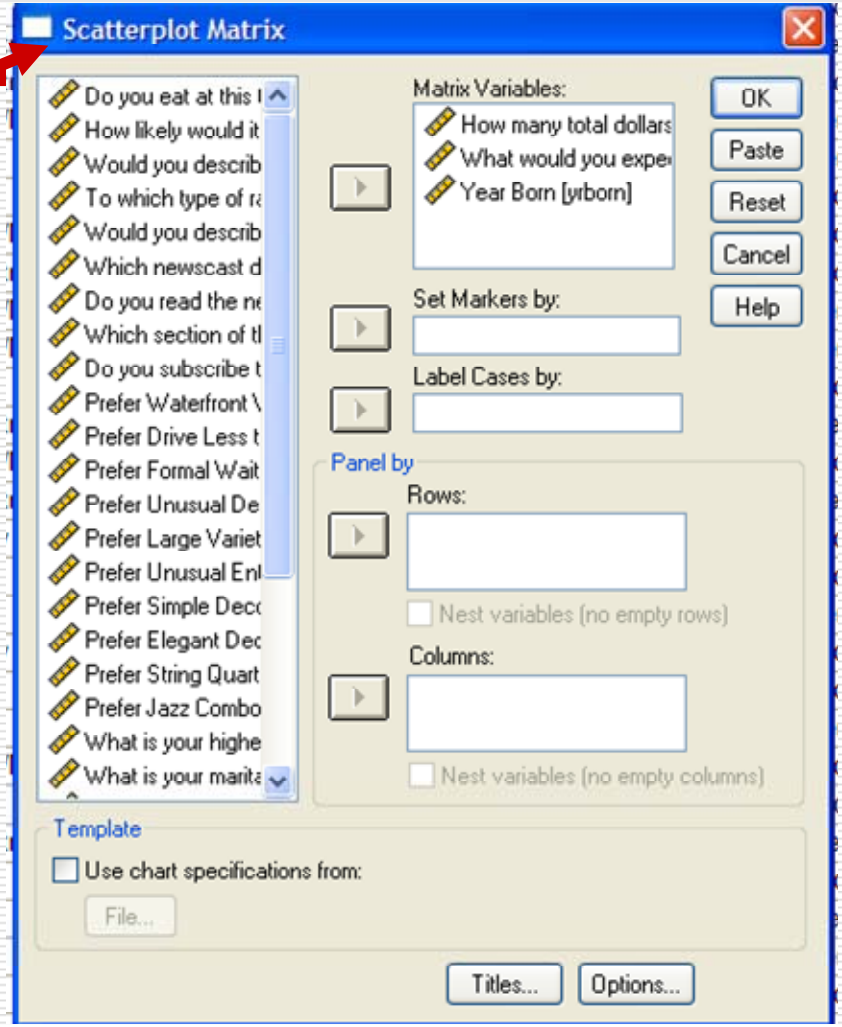


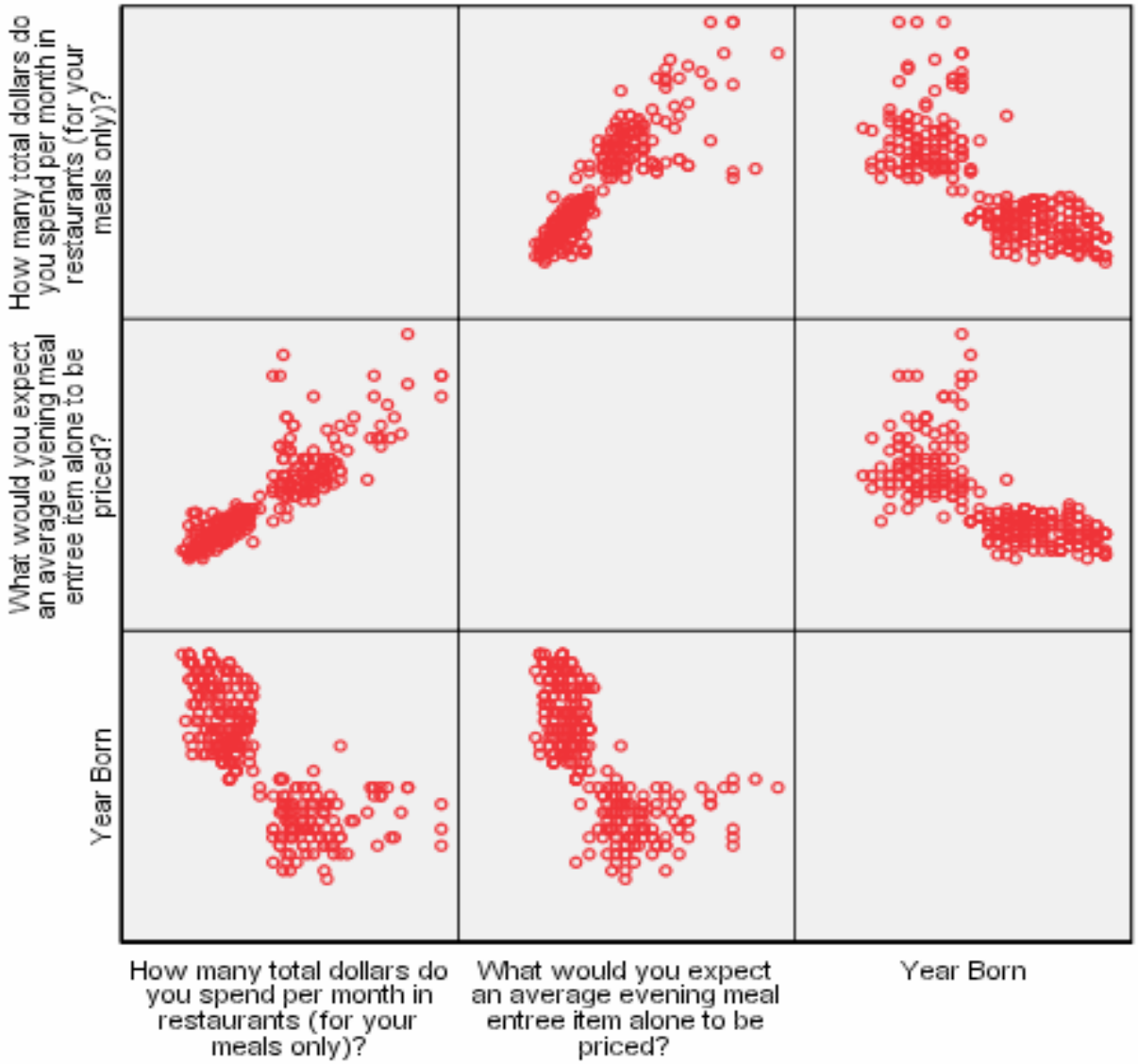
Υπάρχει ισχυρή συσχέτιση

Φαίνεται ομαδοποίηση ως προς την περιοχή

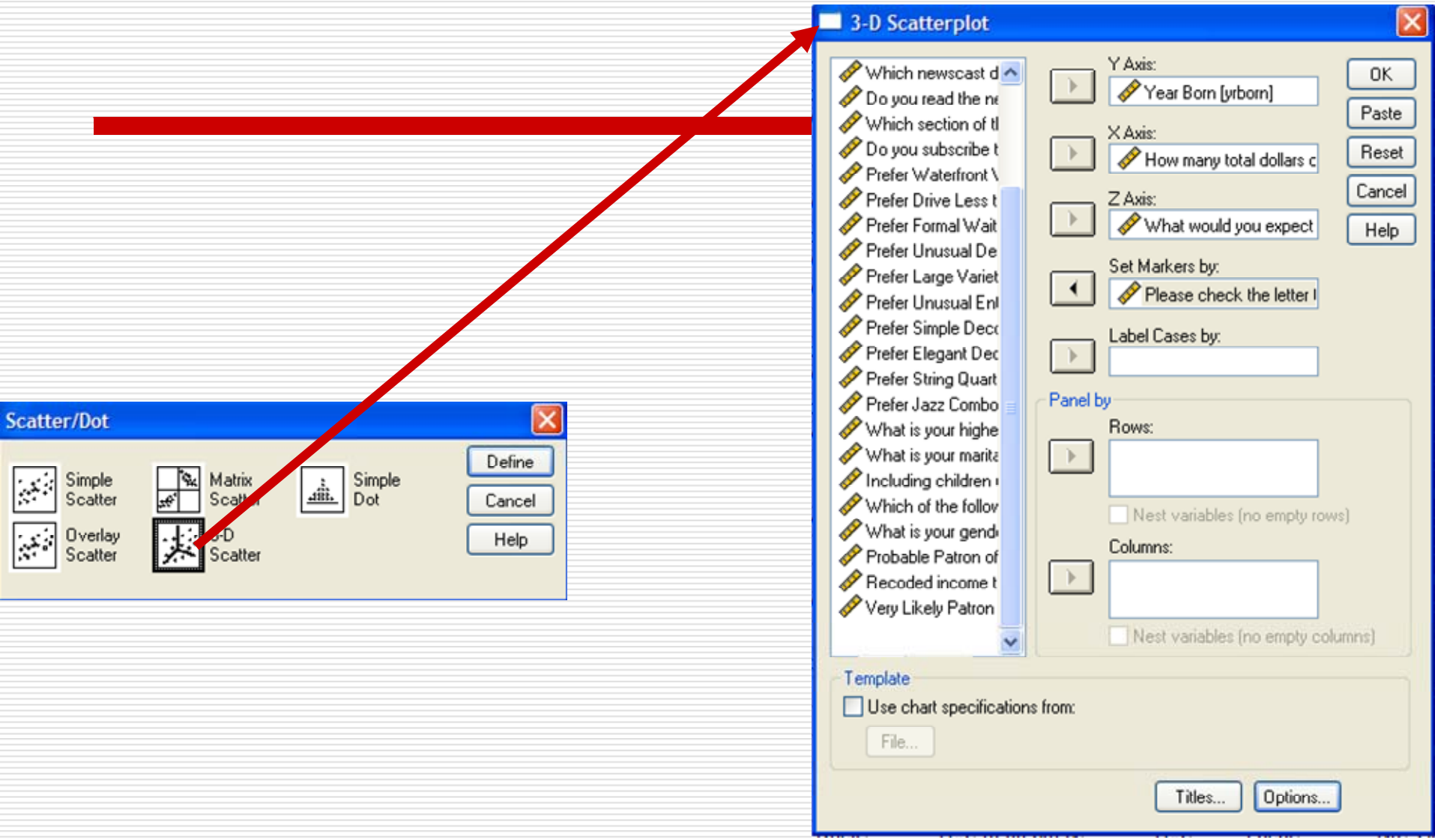
Πίνακας διαγραμμάτων διασποράς για περισσότερες μεταβλητές

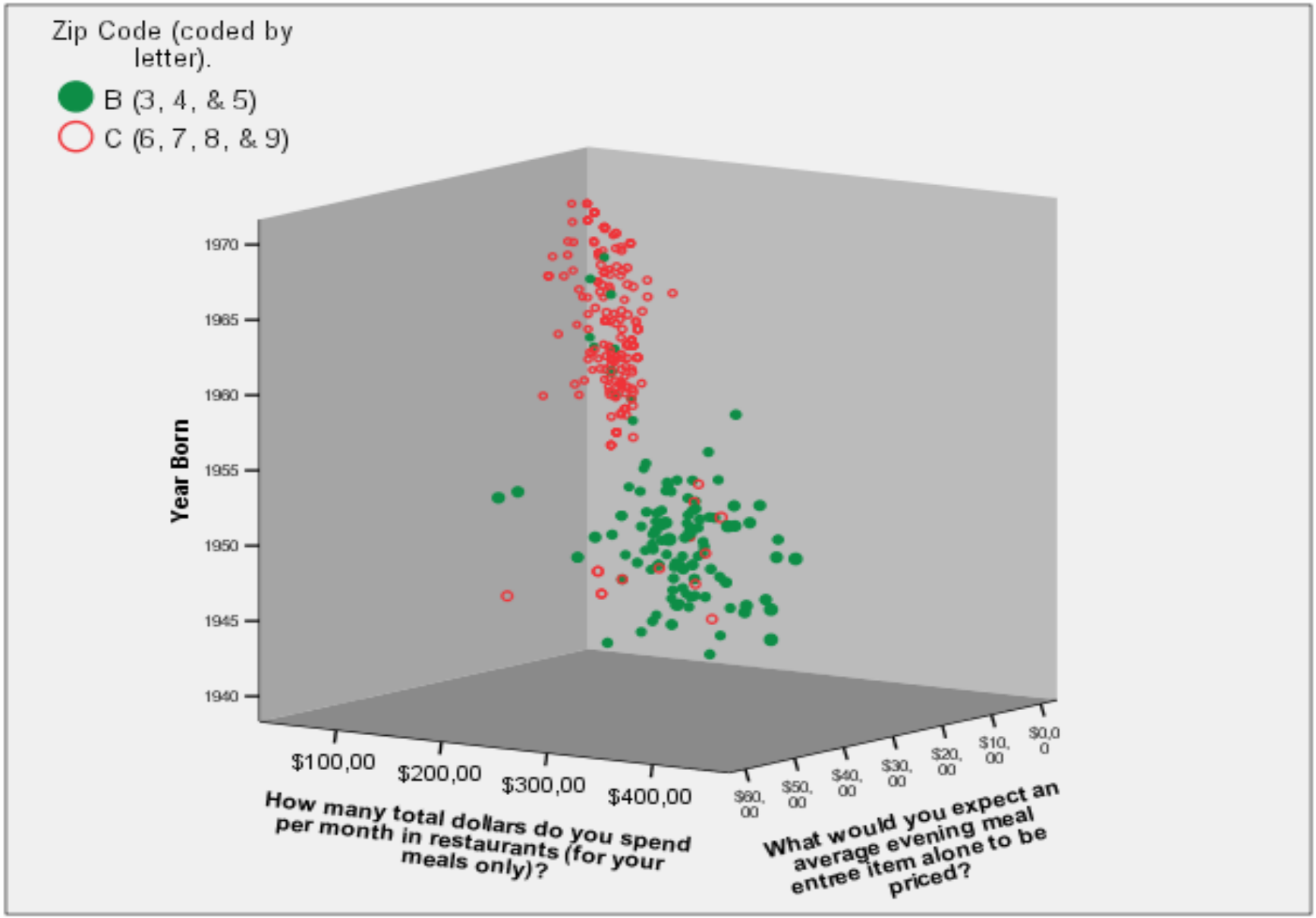




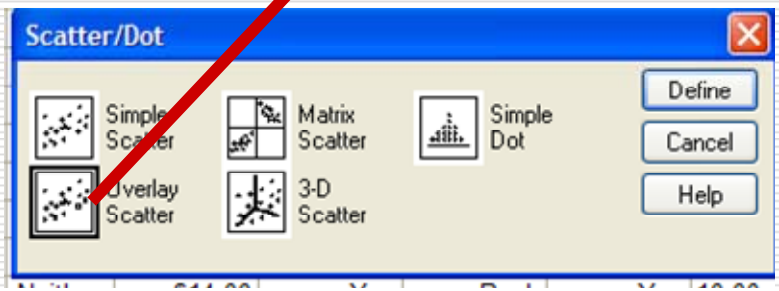

How many total dollars do you spend per month in restaurants (for your meals only)?
What would you expect an average evening meal entree item alone to be priced?
Year Born
How many total dollars do you spend per month in restaurants (for your meals only)?
What would you expect an average evening meal entree item alone to be priced?
Year Born

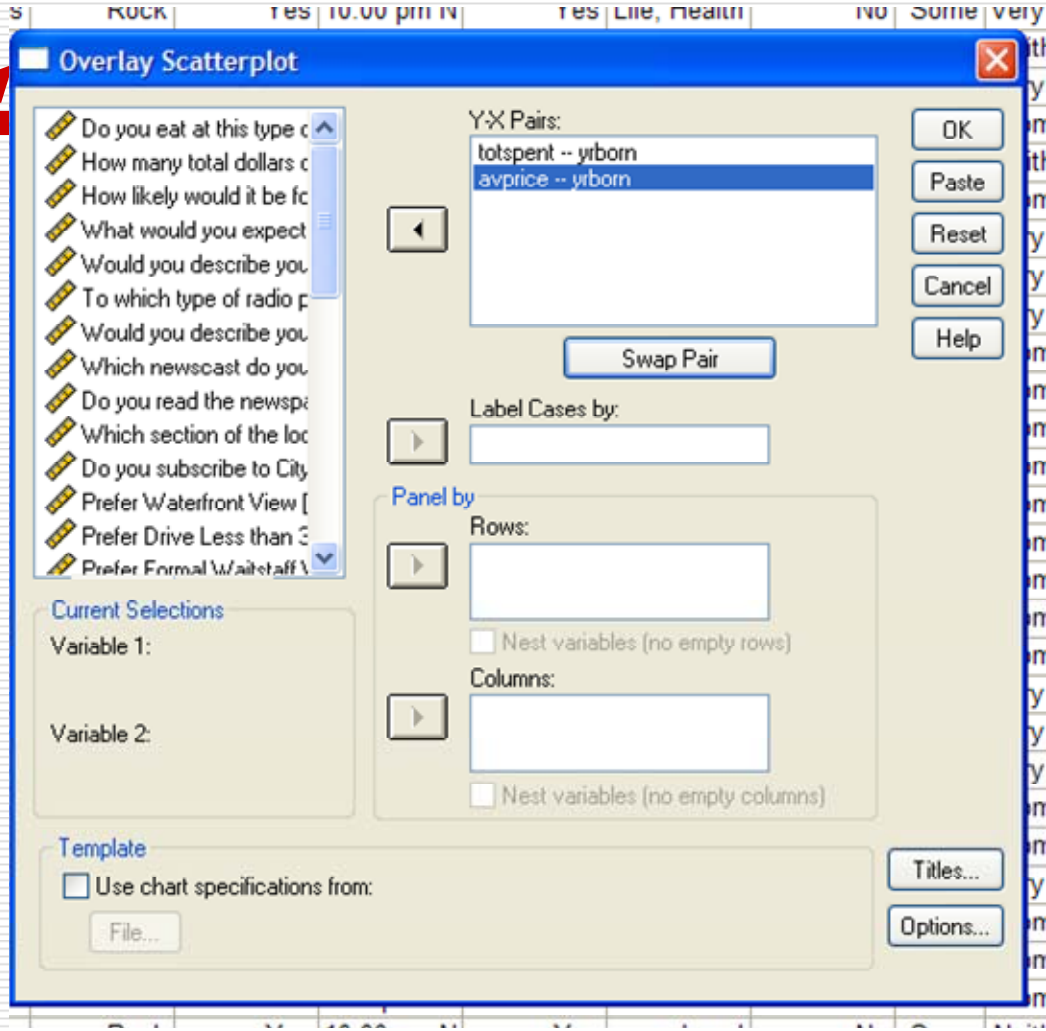
Τρισδιάστατο διάγραμμα διασποράς

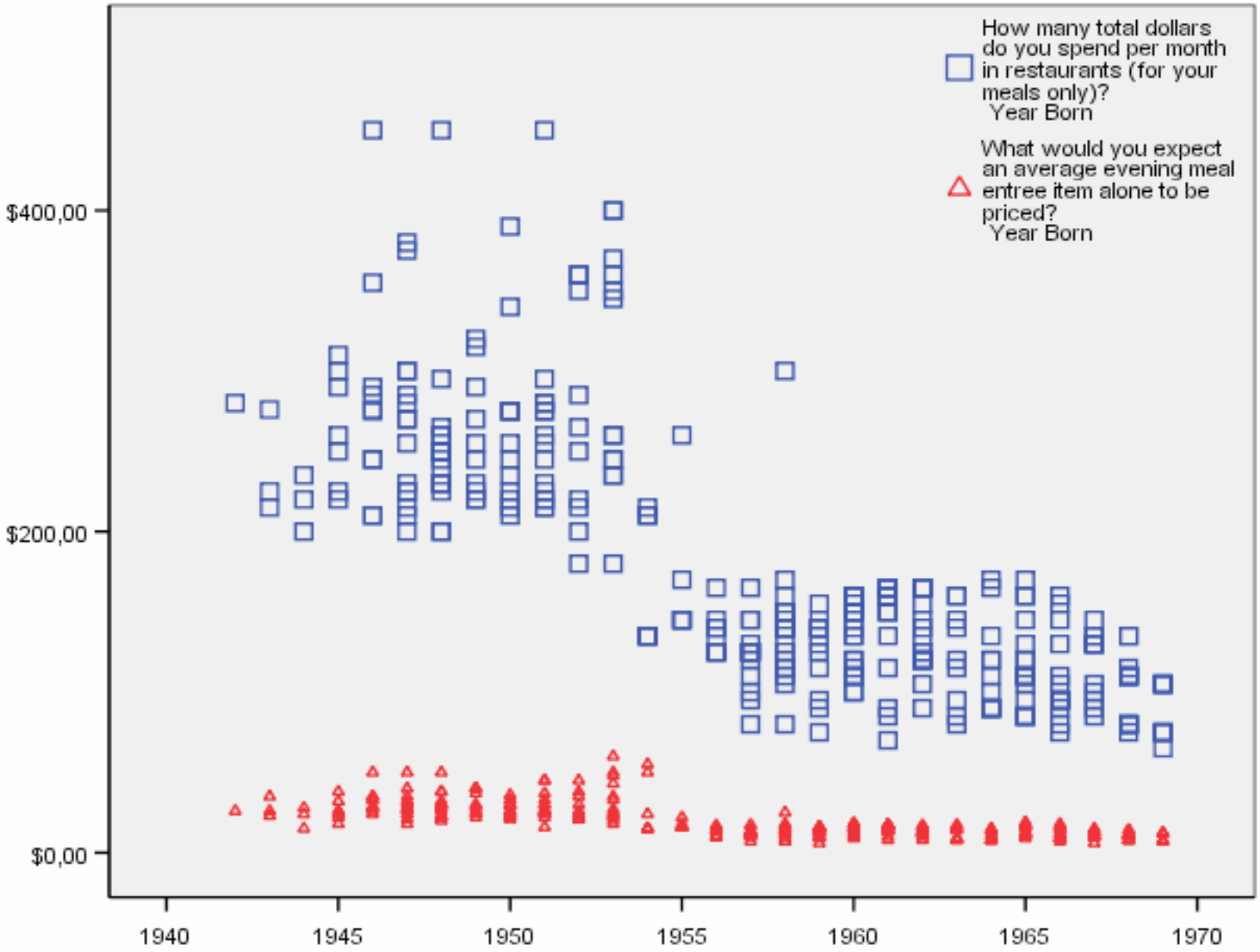



Zip Code (coded by letter).
B (3, 4, & 5)
C (6, 7, 8, & 9)
Year Born
1970
1965
1960
1955
1950
1945
1940
$100,00
$200,00
$300,00
$400,00
How many total dollars do you spend per month in restaurants (for your meals only)?
$60,00
$50,00
$40,00
$30,00
$20,00
$10,00
$0,00
What would you expect an average evening meal entree item alone to be priced?

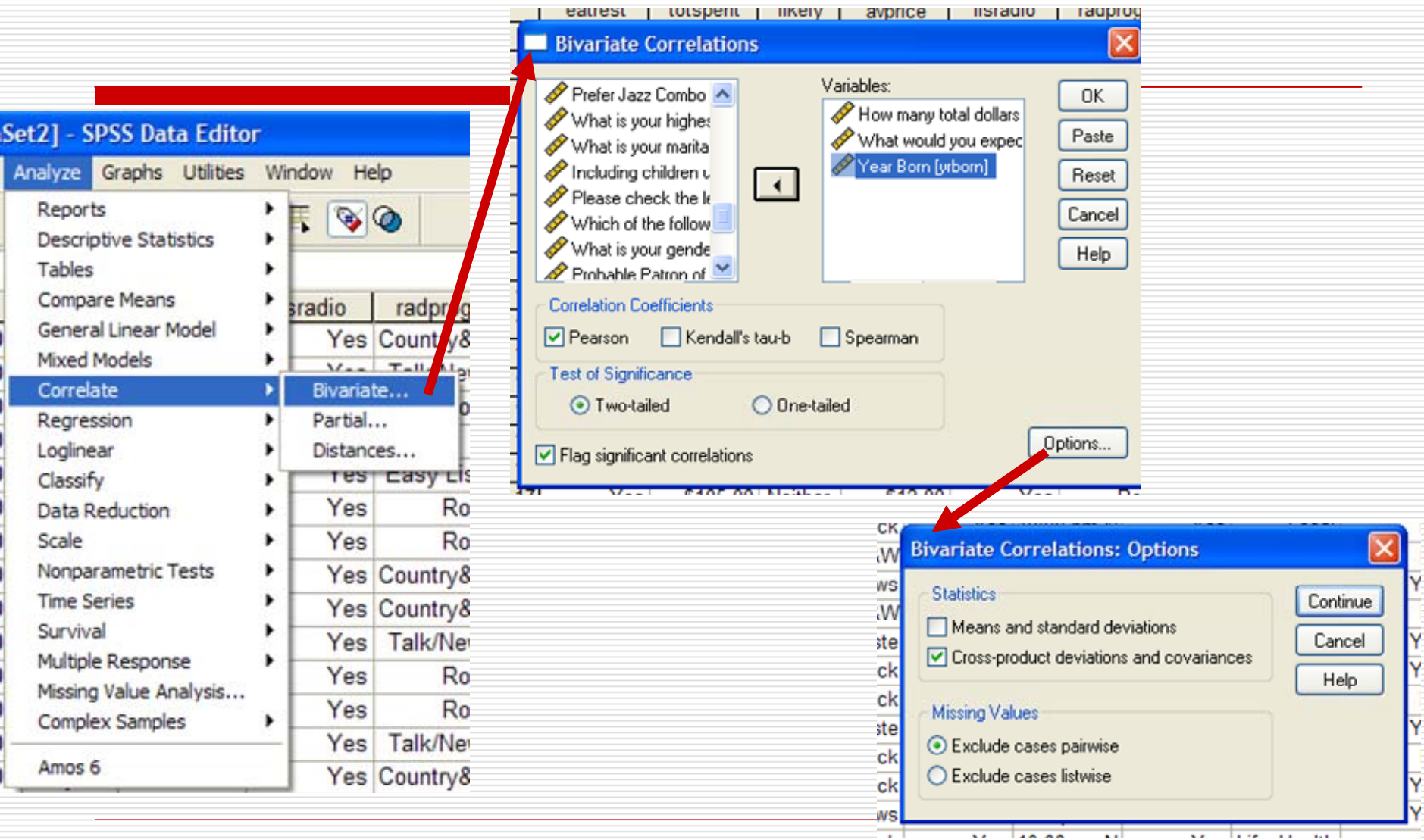
Επικαλυπτόμενα διαγράμματα διασποράς






How many total dollars do you spend per month in restaurants (for your meals only)?
Year Born
What would you expect an average evening meal entree item alone to be priced?
Year Born
$400,00
$200,00
$0,00
1940
1945
1950
1955
1960
1965
1970

Συσχέτιση δύο μεταβλητών (Bivariate correlation)



Θετική συσχέτιση πολύ σημαντική

Αρνητική συσχέτιση πολύ σημαντική

Correlations

		How many total dollars do you spend per month in restaurants (for your meals only)?	What would you expect an average evening meal entree item alone to be priced?	Year Born
How many total dollars do you spend per month in restaurants (for your meals only)?	Pearson Correlation	1	,878**	-,431**
	Sig. (2-tailed)		,000	,000
	Sum of Squares and Cross-products	3429187,898	230222,329	-151651,713
	Covariance	8594,456	679,122	-380,079
	N	400	340	400
What would you expect an average evening meal entree item alone to be priced?	Pearson Correlation	,878**	1	-,722**
	Sig. (2-tailed)	,000		,000
	Sum of Squares and Cross-products	230222,329	32742,776	-17239,071
	Covariance	679,122	96,586	-50,853
	N	340	340	340
Year Born	Pearson Correlation	-,431**	-,722**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	-151651,713	-17239,071	36129,438
	Covariance	-380,079	-50,853	90,550
	N	400	340	400

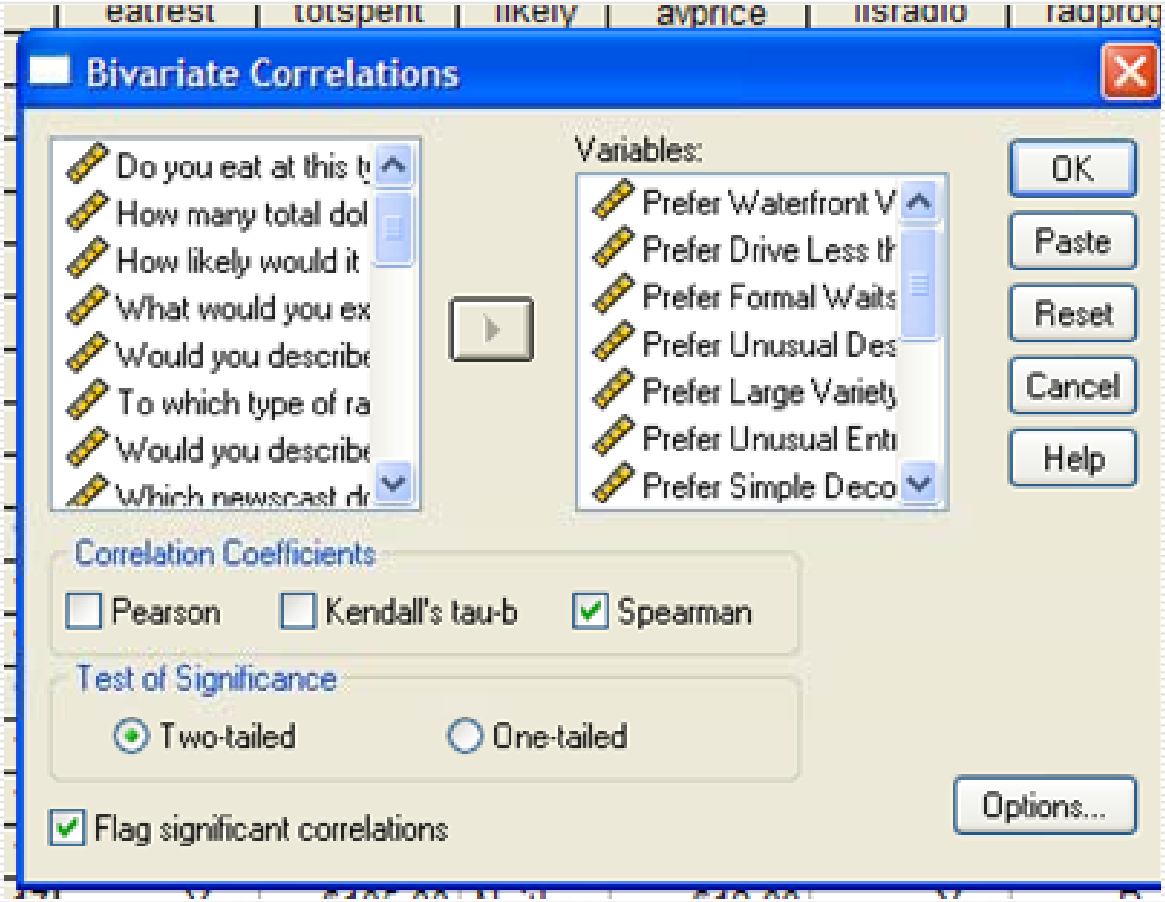
**. Correlation is significant at the 0.01 level (2-tailed).

Ο συντελεστής Pearson κυρίως για συνεχή, κανονικά κατανεμημένα δεδομένα

Ο συντελεστής συσχέτισης του Spearman (Spearman's correlation coefficient)

- Μη-παραμετρικό στατιστικό μέτρο
- Τα δεδομένα δεν είναι ανάγκη να είναι κανονικά ούτε συνεχή
- Βασίζεται σε διάταξη των δεδομένων (ranking) και υπολογισμό του συντελεστή του Pearson στις διατάξεις (ranks)
- Ιδανικό για μεταβλητές διάταξης (ordinal)

Εφαρμογή σε ερωτήσεις αξιολόγησης με απαντήσεις 1-5



Correlations

			Prefer Waterfront View	Prefer Drive Less than 30 Minutes	Prefer Formal Waitstaff Wearing Tuxedos	Prefer Unusual Desserts	Prefer Large Variety of Entrees	Prefer Unusual Entrees	Prefer Simple Decor	Prefer Elegant Decor	Prefer String Quartet	Prefer Jazz Combo
Spearman's rho	Prefer Waterfront View	Correlation Coefficien	1,000	-,723**	-,618**	-,657**	-,683**	-,620**	,597**	-,642**	-,676**	,597**
		Sig. (2-tailed)	.	,000	,000	,000	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Drive Less than 30 Minutes	Correlation Coefficien	-,723**	1,000	,663**	,605**	,724**	,634**	-,663**	,711**	,690**	-,499**
		Sig. (2-tailed)	,000	.	,000	,000	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Formal Waitstaff Wearing Tuxedos	Correlation Coefficien	-,618**	,663**	1,000	,757**	,700**	,745**	-,725**	,776**	,666**	-,481**
		Sig. (2-tailed)	,000	,000	.	,000	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Unusual Desserts	Correlation Coefficien	-,657**	,605**	,757**	1,000	,682**	,746**	-,729**	,726**	,730**	-,485**
		Sig. (2-tailed)	,000	,000	,000	.	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Large Variety of Entrees	Correlation Coefficien	-,683**	,724**	,700**	,682**	1,000	,704**	-,669**	,657**	,595**	-,440**
		Sig. (2-tailed)	,000	,000	,000	,000	.	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Unusual Entrees	Correlation Coefficien	-,620**	,634**	,745**	,746**	,704**	1,000	-,789**	,712**	,722**	-,490**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	.	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Simple Decor	Correlation Coefficien	,597**	-,663**	-,725**	-,729**	-,669**	-,789**	1,000	-,767**	-,723**	,490**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	.	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Elegant Decor	Correlation Coefficien	-,642**	,711**	,776**	,726**	,657**	,712**	-,767**	1,000	,721**	-,492**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	.	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer String Quartet	Correlation Coefficien	-,676**	,690**	,666**	,730**	,595**	,722**	-,723**	,721**	1,000	-,584**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	.	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Jazz Combo	Correlation Coefficien	,597**	-,499**	-,481**	-,485**	-,440**	-,490**	,490**	-,492**	-,584**	1,000
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000	.
		N	400	400	400	400	400	400	400	400	400	400

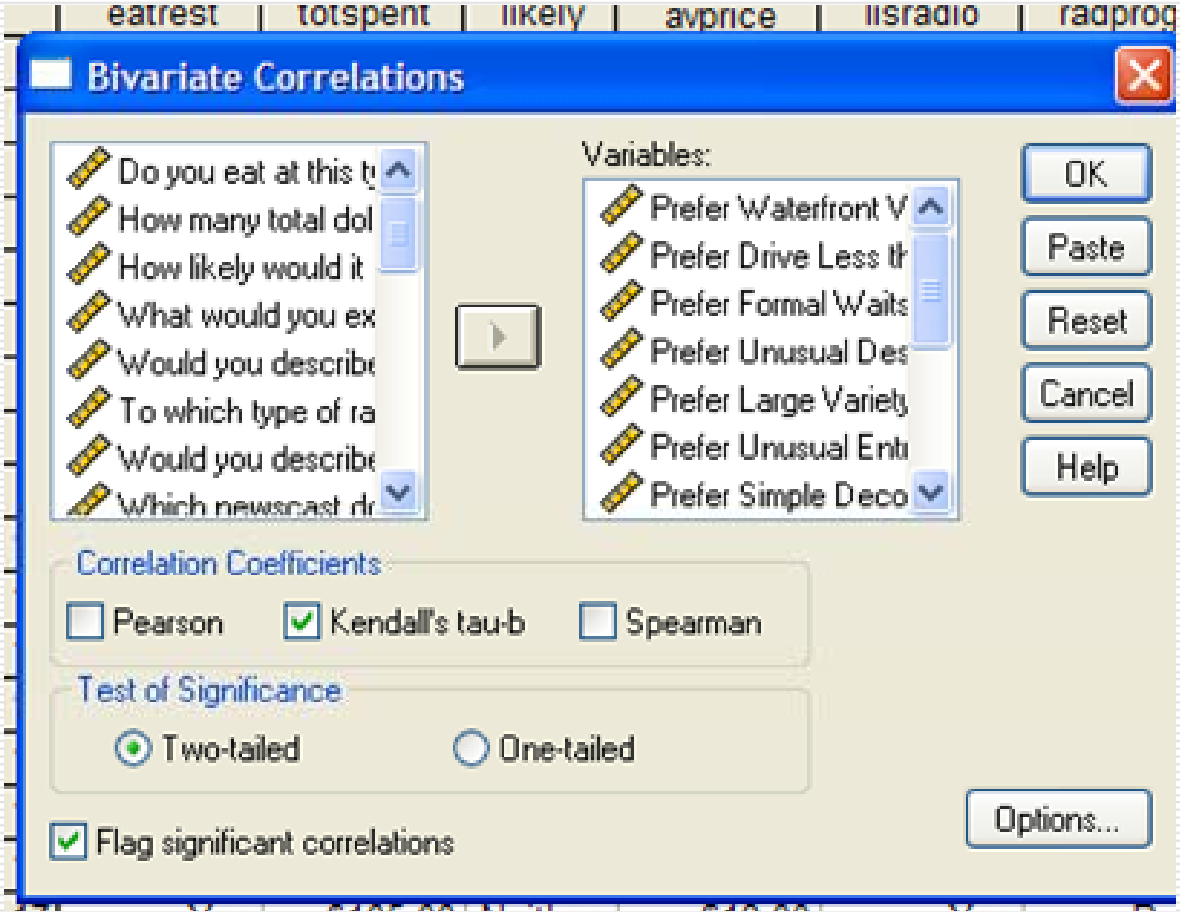
**. Correlation is significant at the 0.01 level (2-tailed).

Υπάρχουν πολύ ισχυρές συσχετίσεις (θετικές και αρνητικές)

Το τ του Kendall (Kendall's tau)

- Μη παραμετρικός συντελεστής
- Για μικρά σύνολα δεδομένων με πολλές «ισοπαλίες» στα ranks
- Θεωρείται καλύτερος εκτιμητής της συσχέτισης που υπάρχει στον πληθυσμό
- Γενικά μικρότερες συσχετίσεις από του Spearman

Εφαρμογές σε μεταβλητές διάταξεις



Correlations

			Prefer Waterfront View	Prefer Drive Less than 30 Minutes	Prefer Formal Waitstaff Wearing Tuxedos	Prefer Unusual Desserts	Prefer Large Variety of Entrees	Prefer Unusual Entrees	Prefer Simple Decor	Prefer Elegant Decor	Prefer String Quartet	Prefer Jazz Combo
Kendall's tau_b	Prefer Waterfront View	Correlation Coefficient	1,000	-,615**	-,482**	-,544**	-,555**	-,487**	,467**	-,507**	-,562**	,476**
		Sig. (2-tailed)	.	,000	,000	,000	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Drive Less than 30 Minutes	Correlation Coefficient	-,615**	1,000	,541**	,477**	,610**	,499**	-,542**	,600**	,574**	-,406**
		Sig. (2-tailed)	,000	.	,000	,000	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Formal Waitstaff Wearing Tuxedos	Correlation Coefficient	-,482**	,541**	1,000	,660**	,594**	,647**	-,616**	,696**	,538**	-,412**
		Sig. (2-tailed)	,000	,000	.	,000	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Unusual Desserts	Correlation Coefficient	-,544**	,477**	,660**	1,000	,558**	,641**	-,640**	,614**	,624**	-,414**
		Sig. (2-tailed)	,000	,000	,000	.	,000	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Large Variety of Entrees	Correlation Coefficient	-,555**	,610**	,594**	,558**	1,000	,596**	-,541**	,548**	,455**	-,351**
		Sig. (2-tailed)	,000	,000	,000	,000	.	,000	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Unusual Entrees	Correlation Coefficient	-,487**	,499**	,647**	,641**	,596**	1,000	-,694**	,606**	,602**	-,412**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	.	,000	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Simple Decor	Correlation Coefficient	,467**	-,542**	-,616**	-,640**	-,541**	-,694**	1,000	-,663**	-,613**	,423**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	.	,000	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Elegant Decor	Correlation Coefficient	-,507**	,600**	,696**	,614**	,548**	,606**	-,663**	1,000	,611**	-,417**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	.	,000	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer String Quartet	Correlation Coefficient	-,562**	,574**	,538**	,624**	,455**	,602**	-,613**	,611**	1,000	-,495**
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	.	,000
		N	400	400	400	400	400	400	400	400	400	400
	Prefer Jazz Combo	Correlation Coefficient	,476**	-,406**	-,412**	-,414**	-,351**	-,412**	,423**	-,417**	-,495**	1,000
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000	.
		N	400	400	400	400	400	400	400	400	400	400

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

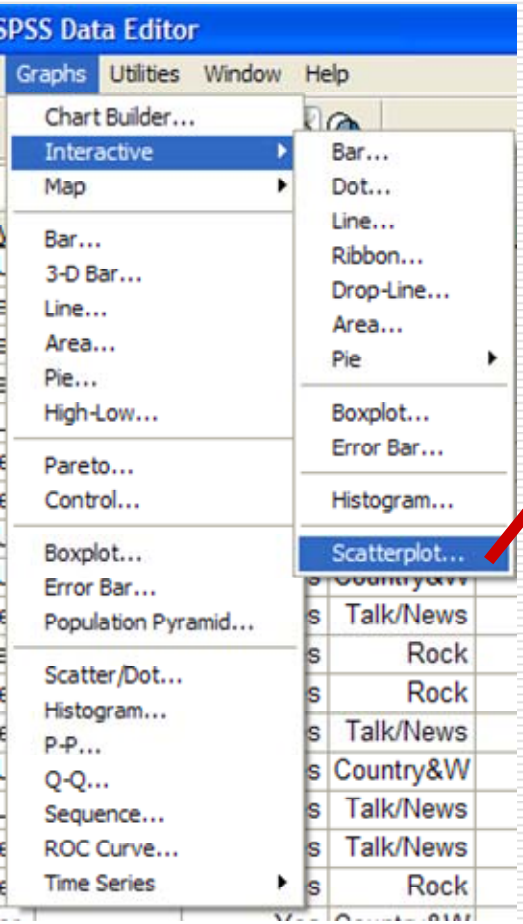
Υπάρχουν πολύ ισχυρές συσχετίσεις (θετικές και αρνητικές)

Μοντελοποίηση της συσχέτισης

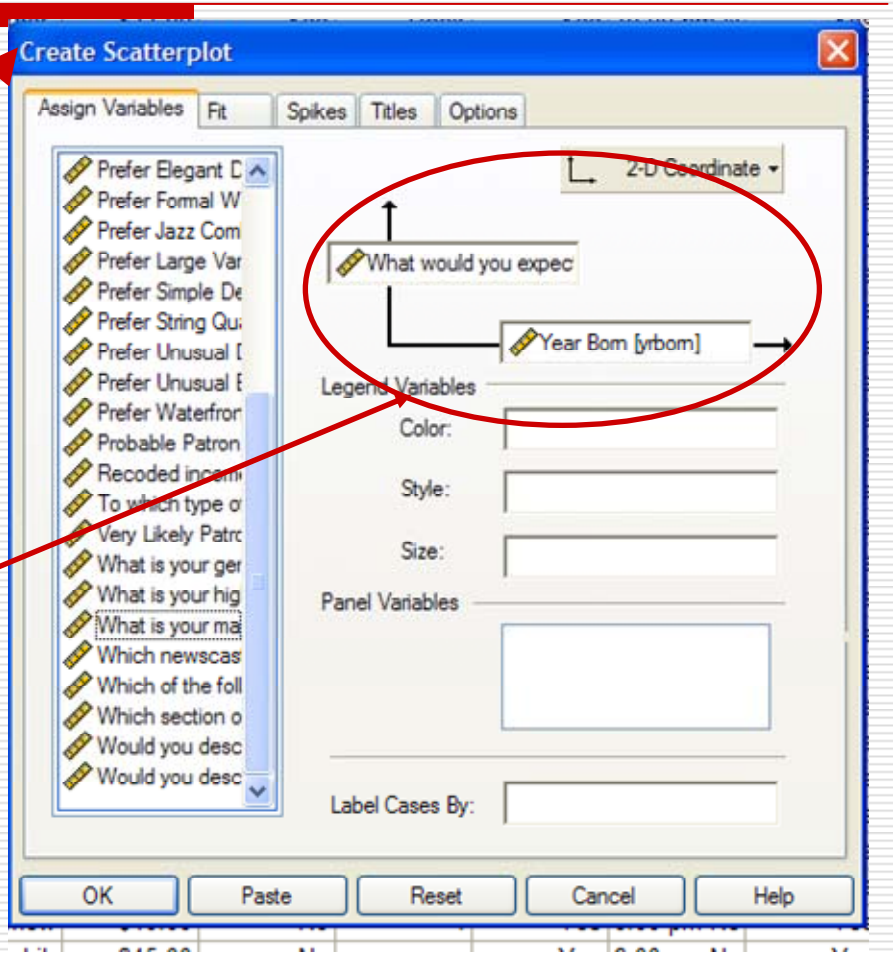
- Εύρεση εξίσωσης που περιγράφει τη σχέση δύο μεταβλητών
- Υπάρχει γενική μεθοδολογία κατασκευής και ελέγχου ενός μοντέλου: **Ανάλυση παλινδρόμησης** (*regression analysis*)
- Η μια μεταβλητή (y) θεωρείται **εξαρτημένη** (*dependent*) από την άλλη (x) η οποία ονομάζεται **ανεξάρτητη** (*independent*)
- Η γενική μεθοδολογία στο επόμενο...

Εύρεση απλού μοντέλου

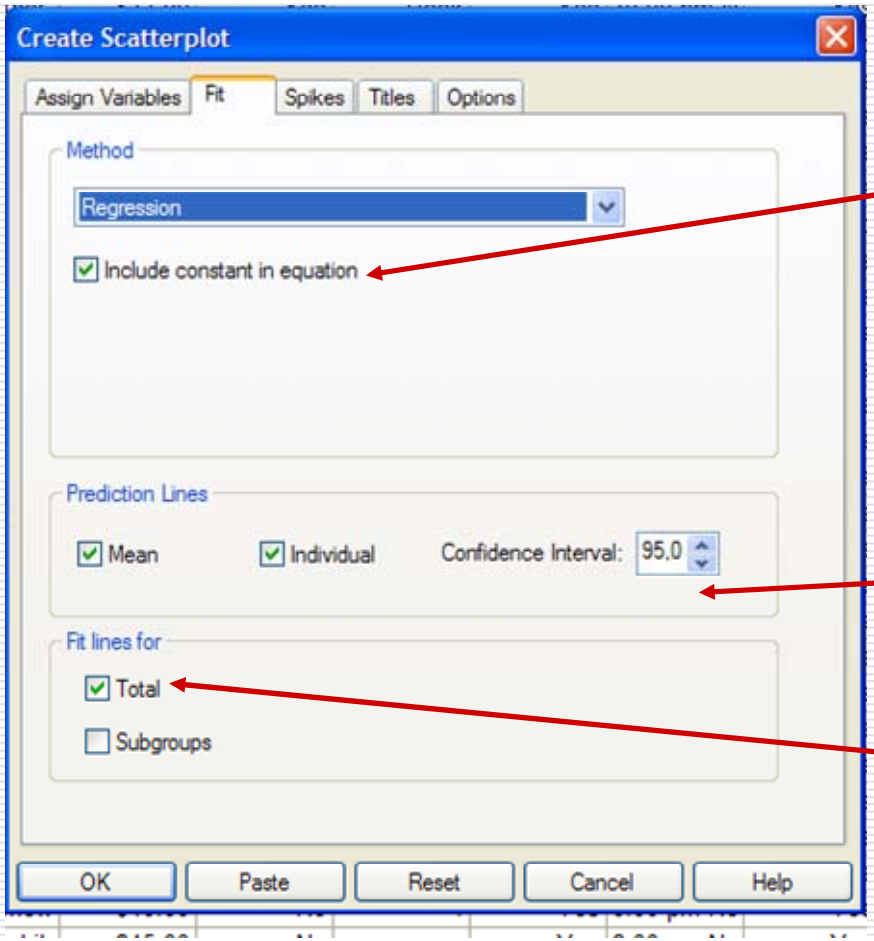
(Παράδειγμα: εξαρτημένη avprice, ανεξάρτητη yrborn)



Ορισμός μεταβλητών



Προσαρμογή (fitting) ευθείας παλινδρόμησης



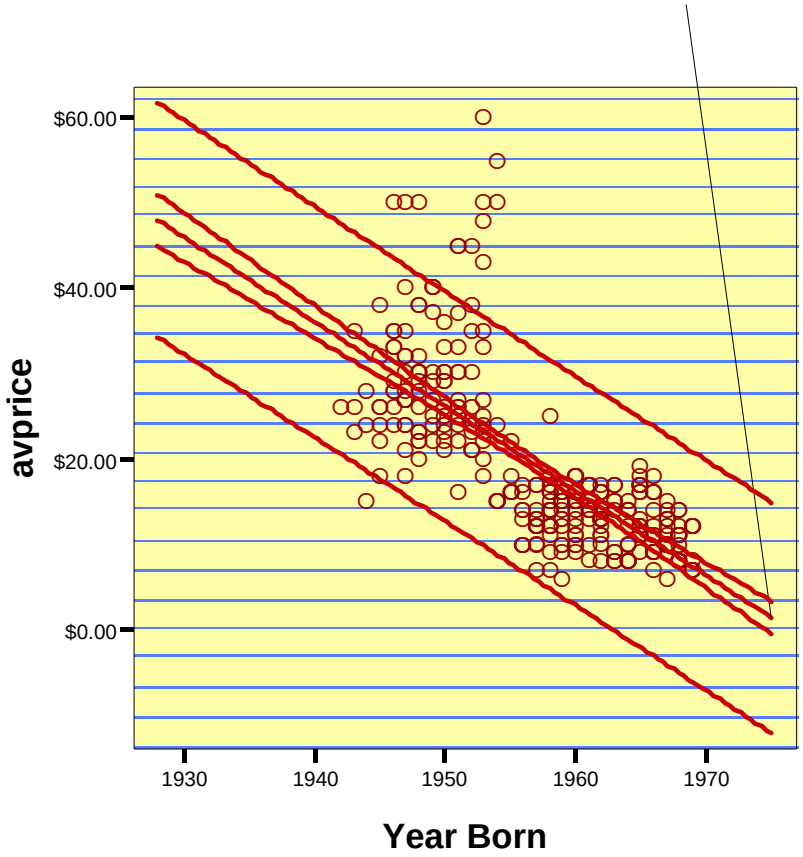
y=a+bx

Ορισμός διαστημάτων εμπιστοσύνης

Μια ευθεία για όλα τα δεδομένα

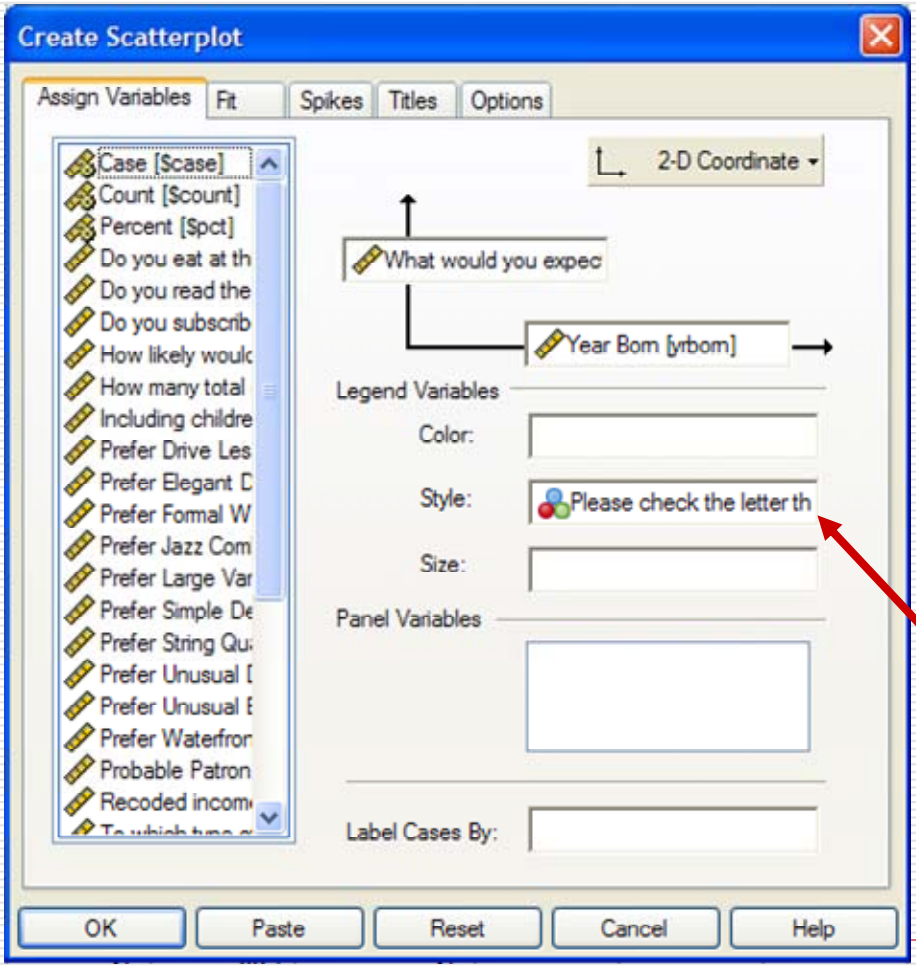
What would you expect an average evening meal entree item alone to be priced? = 1955,87 + -0,99 * yrborn
R-Square = 0,52

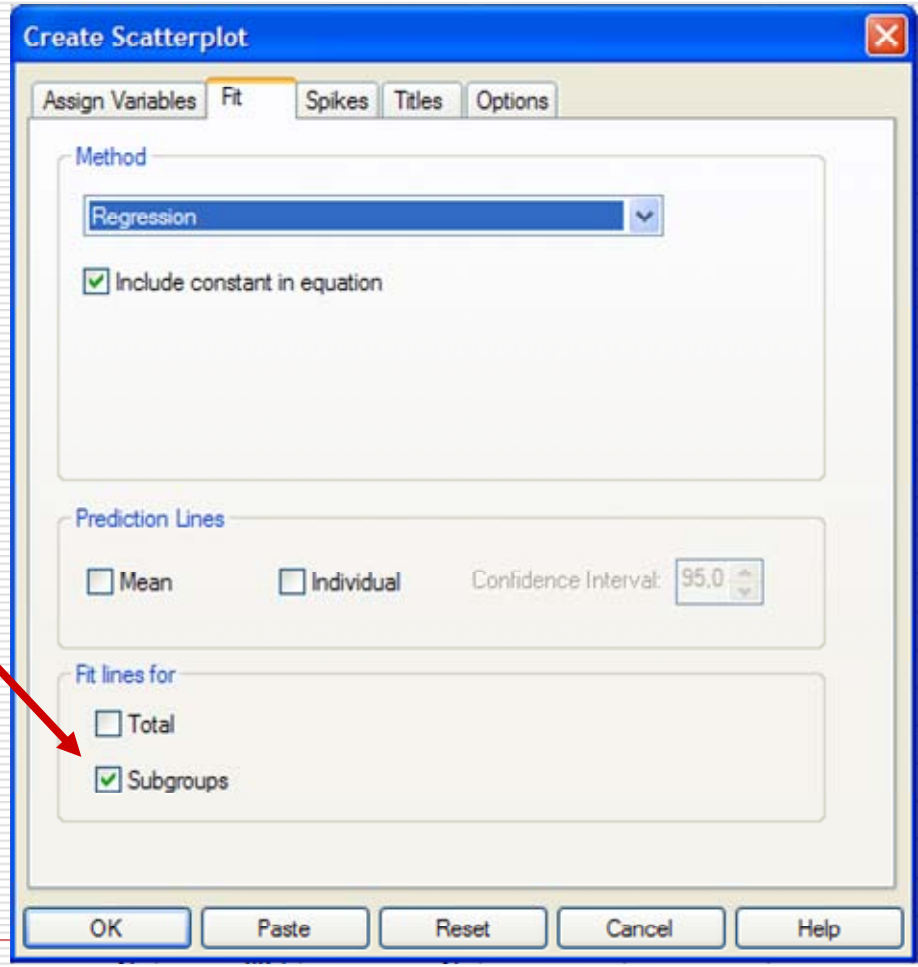
Το τετράγωνο του συντελεστή Pearson δείχνει το ποσοστό μεταβλητότητας της avprice που εξηγείται από την yrborn (52%)



Linear Regression with
95,00% Mean Prediction Interval and
95,00% Individual Prediction Interval

Ορισμός υποομάδων – διαφορετική ευθεία για κάθε μια (π.χ. ως προς τον κωδικό περιοχής)

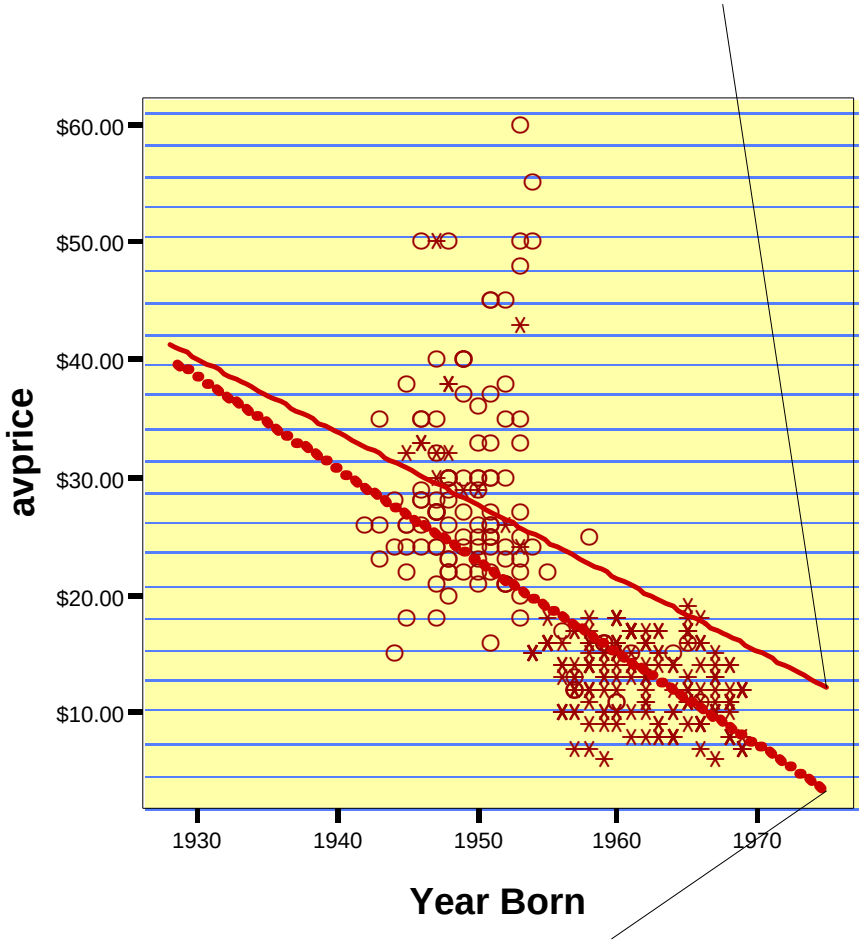




What would you expect an average evening meal entree item alone to be priced? = 1237,09 + -0,62 * yrborn
R-Square = 0,09

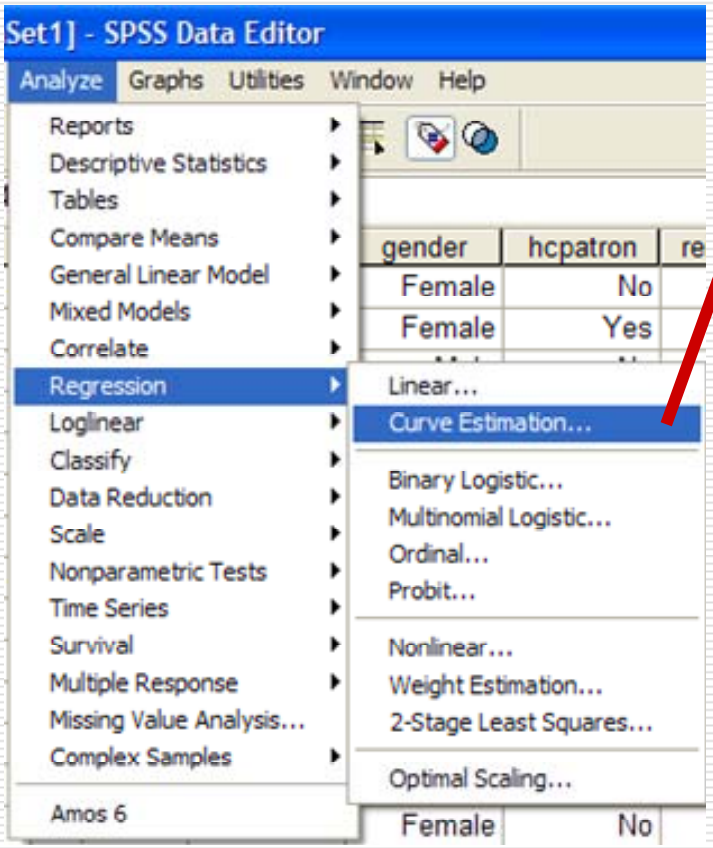
Zip Code (coded by letter).
B (3, 4, & 5)
C (6, 7, 8, & 9)

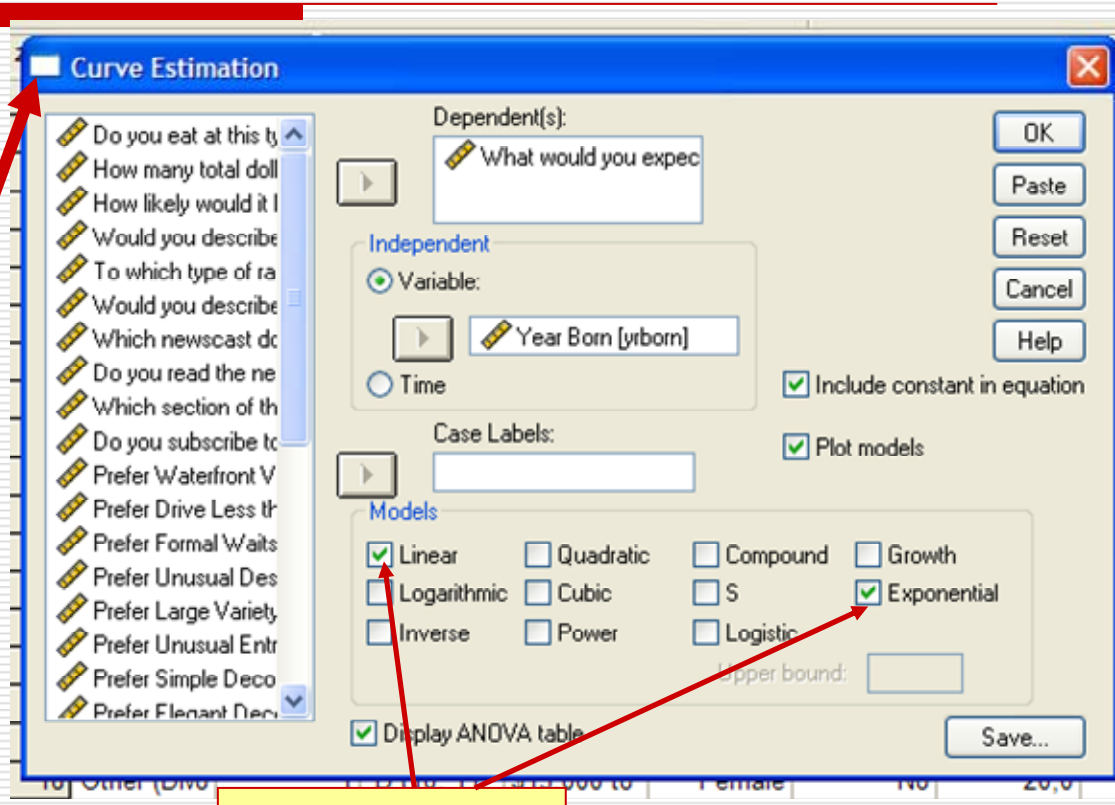
Μόνο στην περιοχή C η ευθεία είναι σχετικά καλή



What would you expect an average evening meal entree item alone to be priced? = 1548,26 + -0,78 * yrborn
R-Square = 0,44

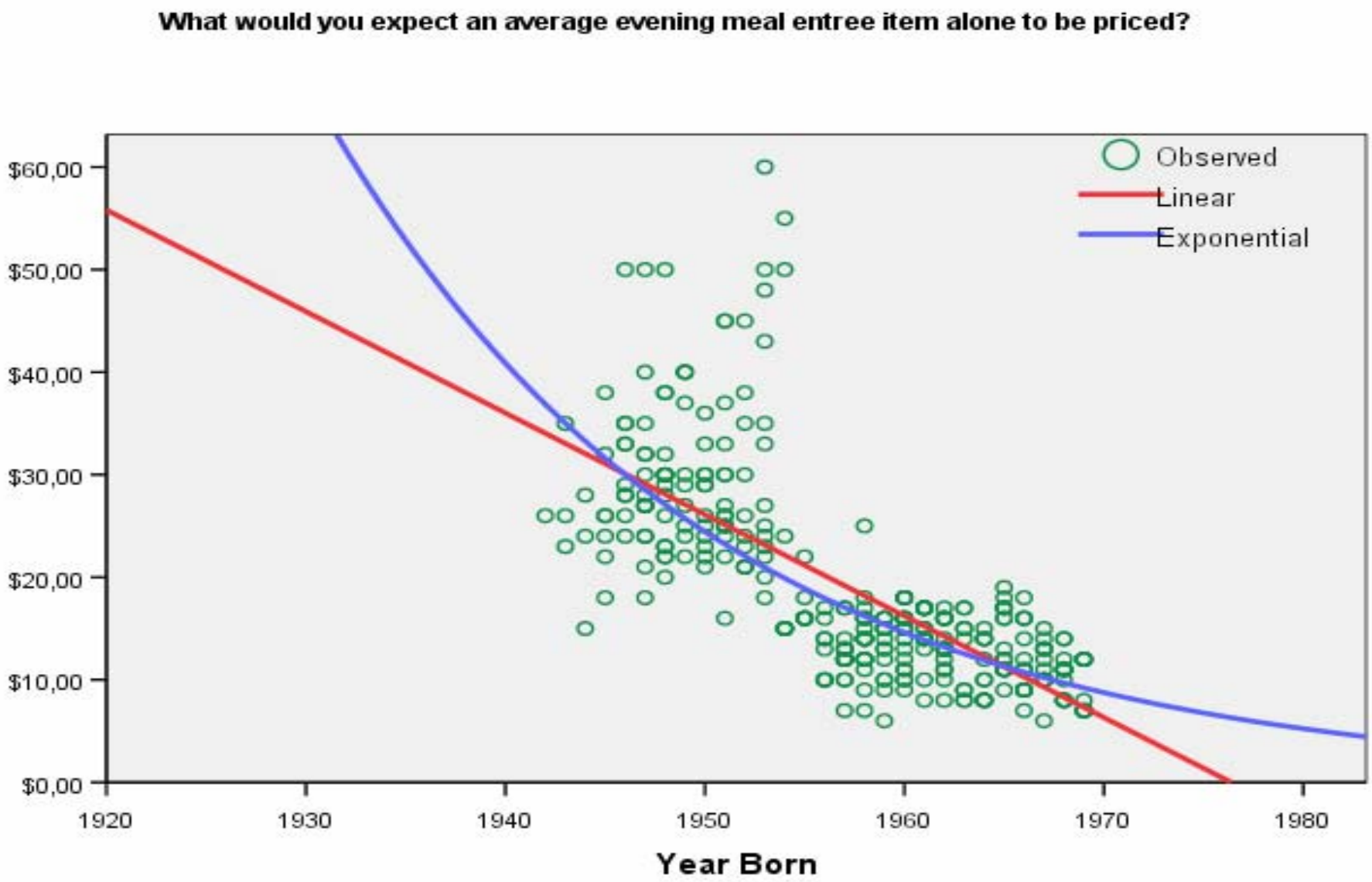
Κατασκευή και άλλων μοντέλων





Γραμμικό και εκθετικό μοντέλο μαζί

Αποτελέσματα – σύγκριση μοντέλων



Γραμμικό μοντέλο
$y=1955.87-0.990*x$

Εκθετικό μοντέλο
$y=(8.2*10^{44})*\exp(-0.051*x)$
ή
$\ln y=103.42-0.051*x$

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,722	,521	,520	6,812

The independent variable is Year Born.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	17060,081	1	17060,081	367,686	,000
Residual	15682,696	338	46,399		
Total	32742,776	339			

The independent variable is Year Born.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Year Born	-,990	,052	-,722	-19,175	,000
(Constant)	1955,868	101,019		19,361	,000

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,775	,600	,599	,301

The independent variable is Year Born.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	46,010	1	46,010	507,941	,000
Residual	30,616	338	,091		
Total	76,626	339			

The independent variable is Year Born.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Year Born	-,051	,002	-,775	-22,538	,000
(Constant)	8,2E+044	3,6E+045		,224	,823

The dependent variable is ln(What would you expect an average evening meal entree item alone to be priced?).

?

Συμπεράσματα

- Οι συντελεστές συσχέτισης και τα διαγράμματα διασποράς μας δείχνουν το μέγεθος και τη φύση της συσχέτισης
- Η μοντελοποίηση της συσχέτισης δεν είναι απλή. Απαιτούνται έλεγχοι του μοντέλου, δυνατότητα ερμηνείας του, εισαγωγή νέων μεταβλητών κλπ