

Άσκηση 1

Μία αντιπροσωπεία πωλήσεως αυτοκινήτων διαθέτει καταστήματα σε 15 διαφορετικές πόλεις. Ο επόμενος πίνακας δίνει τις πωλήσεις Y_i του τελευταίου μήνα καθώς επίσης και τον πληθυσμό X_{i1} και το οικογενειακό μέσο ετήσιο εισόδημα X_{i2} (σε 1000άδες δραχμές) των κατοίκων κάθε πόλης.

i	Πωλήσεις (Y_i)	Πληθυσμός (X_{i1})	Εισόδημα (X_{i2})
1	162	274	2450
2	120	180	3254
3	223	375	3802
4	131	205	2838
5	67	86	2347
6	169	265	3782
7	81	98	3008
8	192	330	2450
9	116	195	2137
10	55	53	2560
11	252	430	4020
12	232	372	4427
13	144	236	2660
14	103	157	2088
15	212	370	2605

Να γίνει το διάγραμμα διασποράς (scatterplot) μεταξύ των X_1 , Y και X_2 , Y .

1. Να βρεθεί το καλύτερο γραμμικό μοντέλο (με βάση την αρχή ελαχίστων τετραγώνων) για την πρόβλεψη των πωλήσεων.

α. Με τη χρήση του πληθυσμού X_1 , δηλ. $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

β. Με τη χρήση του εισοδήματος X_2 , δηλ. $Y = \beta_0 + \beta_1 X_2 + \varepsilon$

Ποια από τις δύο προβλέπουσες μεταβλητές (X_1 ή X_2) είναι καλύτερη; Μετονομάστε την σε X .

2. Να κάνετε μελέτη του μοντέλου $Y = \beta_0 + \beta_1 X + \varepsilon$. Συγκεκριμένα:

α. Κάνετε τους ελέγχους σε ε.σ. $\alpha=5\%$ και $\alpha=1\%$:

$H_0: \beta_1=0$, $H_1: \beta_1 \neq 0$, και $H_0: \beta_1=0$, $H_1: \beta_1 > 0$

$H_1: \beta_0=0$, $H_1: \beta_0 \neq 0$, και $H_0: \beta_0=0$, $H_1: \beta_0 > 0$

$H_0: \beta_0=0.5$, $H_1: \beta_1 \neq 0.5$, και $H_0: \beta_0=0.5$, $H_1: \beta_1 > 0.5$

β. Δώστε δ.ε. με σ.ε. 95% και 99% για τα β_0 , β_1 .

γ. Κάνετε το διάγραμμα διασποράς των δεδομένων (X, Y) μαζί με την προσαρμοσμένη ευθεία και τις ζώνες εμπιστοσύνης (για ατομική και μέση πρόβλεψη) με σ.ε. 95% και 99%.

δ. Πόσες πωλήσεις αναμένονται σε μία πόλη 250.000 χιλιάδων κατοίκων; (να γίνει σημειακή εκτίμηση και να δοθεί το δ.ε. με σ.ε. 95% και 99%). Να δοθεί επίσης το δ.ε. συντελεστού 95% για την ατομική πρόβλεψη.

Ποιος είναι περίπου ο πληθυσμός μιας πόλης όπου τον τελευταίο μήνα πουλήθηκαν 120 μονάδες;

ε. Σε ποίο σημείο X βρίσκουμε το μικρότερο δ.ε. για την μέση πρόβλεψη και γιατί; ποιο είναι εδώ αυτό το δ.ε.;

στ. Να κατασκευάσετε τον πίνακα ανάλυσης διασποράς (ANOVA) και να κάνετε τον έλεγχο του μοντέλου μέσω του F-τεστ.

Απαντήσεις

Αρχικά εισάγουμε τα δεδομένα στο SPSS και μετονομάζουμε τις στήλες σε Y , X_1 , X_2 . Το διδιάστατο διάγραμμα διασποράς μπορεί να γίνει χρησιμοποιώντας:

/Graphs/Scatter (simple, Define... Y , X_1 ή Y , X_2) (use Labels 1,2,...), apply changes to graph..

ή

/Graphs/Interactive/Scatterplot (2D, Define Y , X_1 ή Y , X_2),

interact: Insert element(regression Fit,Spikes), Chart manager (Regression, Edit Prediction intervals)

Ανάλογα, το τριδιάστατο διάγραμμα διασποράς μπορεί να γίνει χρησιμοποιώντας:

/Graphs/Scatter (3D, Define... Y , X_1 , X_2), apply changes to graph..

ή

/Graphs/Interactive/Scatterplot (3D, Define Y , X_1 , X_2), περιστροφή

1. Ενεργοποιούμε τη διαδικασία

/Analyze/Regression/Linear, enter variables (Y , X_1),

statistics (Estimates, conf. Intervals, Model Fit)

options (Include constant)

α. Το μοντέλο $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,995 ^a	,991	,990	6,11

a. Predictors: (Constant), X_1

Ο παραπάνω πίνακας περιέχει τις ποσότητες:

R	R^2	Adj. R^2	Std. Error of Est.
$R = \sqrt{1 - \frac{SSE}{SST}}$	$R^2 = 1 - \frac{SSE}{SST}$	$R^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$	$s = \sqrt{\frac{SSE}{n-2}}$

β. Το μοντέλο $Y = \beta_0 + \beta_1 X_2 + \varepsilon$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,639	,409	,363	49,51

a Predictors: (Constant), X_2

Επομένως, η X_1 δίνει το μεγαλύτερο R^2 και την μετονομάζουμε σε X .

2. α. Εκτελούμε παλινδρόμηση με τα X , Y :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,995 ^a	,991	,990	6,11

a. Predictors: (Constant), X_1

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	22,792	3,727		6,115	,000	14,740	30,844
X1	,529	,014	,995	37,844	,000	,499	,559

a. Dependent Variable: Y

Ο παραπάνω πίνακας περιέχει τις ποσότητες:

	B	Std.Error	t	Sig (p-value)	LB, UB
β_0	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	$s(\hat{\beta}_0) = s \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$	$T^* = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)}$	$P(T > T^* \mid T \sim t_{n-2})$	$\hat{\beta}_0 \pm s(\hat{\beta}_0) t_{n-2}(a/2)$
β_1	$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$	$s(\hat{\beta}_1) = \frac{s}{\sqrt{S_{XX}}}$	$T^* = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$	$P(T > T^* \mid T \sim t_{n-2})$	$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_{n-2}(a/2)$

Επειδή το p -value (=Sig.) = 0,000 απορρίπτουμε την $H_0: \beta_1=0$ έναντι της $H_1: \beta_1 \neq 0$ σε ε.σ. 0.05 και 0.01.

Γενικά, για τον έλεγχο της $H_0: \beta_1=0$ ή $\beta_1 \leq 0$ έναντι της $H_1: \beta_1 > 0$ χρειαζόμαστε το αντίστοιχο p -value. Η περιοχή απόρριψης είναι της μορφής $T > c$ οπότε το p -value θα είναι

$$p\text{-value} = P(T > T^* \mid T \sim t_{n-2})$$

αν το T^* είναι θετικό, τότε (λόγω της συμμετρικότητας της κατανομής t)

$$p\text{-value} = P(T > T^* \mid T \sim t_{n-2}) = \frac{1}{2} P(|T| > |T^*| \mid T \sim t_{n-2}) = \frac{1}{2} \text{Sig}$$

αν το T^* είναι αρνητικό, τότε

$$p\text{-value} = P(T > T^* \mid T \sim t_{n-2}) = 1 - \frac{1}{2} P(|T| > |T^*| \mid T \sim t_{n-2}) = 1 - \frac{1}{2} \text{Sig}$$

Αν είχαμε τον έλεγχο $H_0: \beta_1=0$ ή $\beta_1 \geq 0$ έναντι της $H_1: \beta_1 < 0$ τότε η περιοχή απόρριψης είναι της μορφής $T < c$ οπότε αν το T^* είναι θετικό,

$$p\text{-value} = P(T < T^* \mid T \sim t_{n-2}) = 1 - \frac{1}{2} P(|T| > |T^*| \mid T \sim t_{n-2}) = 1 - \frac{1}{2} \text{Sig}$$

ενώ αν το T^* είναι αρνητικό, τότε

$$p\text{-value} = P(T < T^* \mid T \sim t_{n-2}) = \frac{1}{2} P(|T| > |T^*| \mid T \sim t_{n-2}) = \frac{1}{2} \text{Sig}$$

Για τον συγκεκριμένο τώρα έλεγχο της $H_0: \beta_1=0$ έναντι της $H_1: \beta_1 > 0$ χρειαζόμαστε το

$$p\text{-value} = P(T > 37.84 \mid T \sim t_{n-2}) = \frac{1}{2} P(|T| > 37.84 \mid T \sim t_{n-2}) = \frac{1}{2} \text{Sig.} = .000$$

Άρα και σε αυτή την περίπτωση, απορρίπτουμε την $H_0: \beta_1=0$ έναντι της $H_1: \beta_1 > 0$ σε ε.σ. 0.05 και 0.01.

Αν είχαμε τον έλεγχο $H_0: \beta_1=0$ έναντι της $H_1: \beta_1 < 0$ τότε

$$p\text{-value} = P(T < 37.84 \mid T \sim t_{n-2}) = 1 - \frac{1}{2} P(|T| > 37.84 \mid T \sim t_{n-2}) = 1 - .000 = 1$$

και άρα δεχόμαστε ότι $\beta_1=0$ (ή ισοδύναμα $\beta_1 \geq 0$) έναντι της $\beta_1 < 0$.

Ο έλεγχος $H_0: \beta_1=0.5$, $H_1: \beta_1 \neq 0.5$ (και $H_0: \beta_1=0.5$, $H_1: \beta_1 > 0.5$) δεν γίνεται αυτόματα από το πακέτο. Για το λόγο αυτό θα πρέπει να μετασχηματίσουμε τα δεδομένα ώστε να καταλήξουμε σε έλεγχο της μορφής $\beta'_1 = 0$. Συγκεκριμένα λαμβάνοντας $\beta'_1 = \beta_1 - 0.5$ προκύπτει το μοντέλο

$$Y = \beta_0 + \beta_1 X + \varepsilon \Leftrightarrow Y = \beta_0 + (\beta'_1 + 0.5)X + \varepsilon \Leftrightarrow Y - 0.5X = \beta_0 + \beta'_1 X + \varepsilon$$

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	22,792	3,727		6,115	,000
	X1	2,871E-02	,014	,495	2,055	,061

a Dependent Variable: Y1

και επειδή $p\text{-value} = 0.061 > 0.05$ δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. 0.05. Για τον μονό-πλευρο έλεγχο διατηρώντας το ίδιο ε.σ. παίρνουμε το μισό $p\text{-value}$ ($H_1: \beta_1 > 0.5$) ή το 1 μείον το μισό $p\text{-value}$ ($H_1: \beta_1 < 0.5$).

β. Τα δ.ε. 95% για τα β_0, β_1 δίνονται απευθείας στον παραπάνω πίνακα. Για να κατασκευάσουμε δ.ε. 99%, θα πρέπει να χρησιμοποιήσουμε τον παρακάτω τύπο:

$$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_{n-2}(0.01/2) = 0,5287140298374 \pm 0,01397102606677 \cdot t_{13}(0.005)$$

Χρησιμοποιούμε την επιλογή /Transform/compute στον SPSS data editor πληκτρολογώντας

$$0.5287140298374 + 0.01397102606677 * \text{IDF.T}(0.995, 13)$$

σε μία νέα μεταβλητή για το άνω όριο, και

$$0.5287140298374 - 0.01397102606677 * \text{IDF.T}(0.995, 13)$$

για το κάτω όριο. Συγκεκριμένα παίρνουμε το (0.4866, 0.5708).

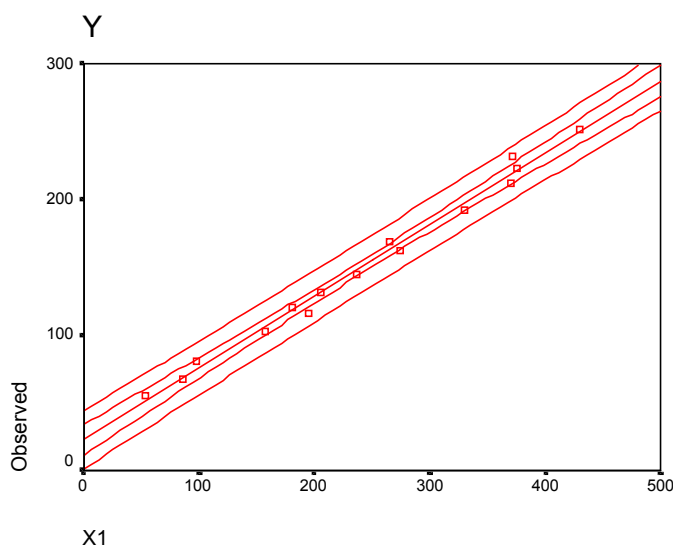
γ. Το διάγραμμα διασποράς μπορεί να γίνει με δύο τρόπους:

1. /Analyze/Regression/Curve estimation (define variables, include constant, linear model)

Στη συνέχεια επεξεργαζόμαστε το σχήμα που κατασκευάστηκε: Στον SPSS Chart Editor

/Gallery/scatter/simple, Replace: Y: Observed, X: X1

/Chart/options/Fit line → Total, Fit options → Linear Regression, Regression Prediction lines (95% και 99%)



Οι ζώνες εμπιστοσύνης είναι: η μία για την μέση πρόβλεψη στο X_0 , $E(\hat{Y}_0) = \beta_0 + \beta_1 X_0$,

$$\hat{Y}_o \pm s(\hat{Y}_o) \cdot t_{n-2}(a/2) = \hat{\beta}_0 + \hat{\beta}_1 X_o \pm s \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{XX}}} \cdot t_{n-2}(a/2)$$

και η άλλη για την ατομική πρόβλεψη στο X_o , $Y_o = \beta_0 + \beta_1 X_o + \varepsilon$

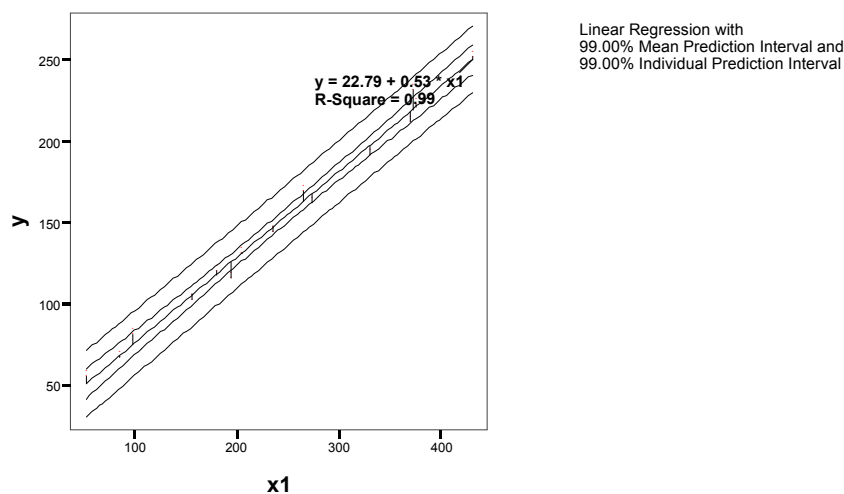
$$\hat{Y}_o \pm s(Y_o) \cdot t_{n-2}(a/2) = \hat{\beta}_0 + \hat{\beta}_1 X_o \pm s \sqrt{1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{XX}}} \cdot t_{n-2}(a/2)$$

Οι τιμές των άκρων των διαστημάτων αυτών μπορούν να εξαχθούν στον SPSS data editor χρησιμοποιώντας την επιλογή **Save** στο μενού **/Analyze/Regression/linear**.

- Εναλλακτικά, θα μπορούσαμε να κατασκευάσουμε το γράφημα χρησιμοποιώντας:

/Graphs/Interactive/Scatterplot (2D, Define Y, X1 ή Y, X2),

interact: Insert element(regression Fit,Spikes), Chart manager (Regression, Edit Prediction intervals)



Με δεξί click στο περιθώριο του (ανοιχτού) σχήματος μπορούμε να κάνουμε διάφορες αλλαγές (π.χ. ChartLooks..).

δ. Οι πωλήσεις που αναμένονται σε μία πόλη με 250 χιλ. κατοίκους θα είναι (εκτίμηση του $E(Y)$ όταν $X = 250$)

$$\hat{Y}_o = \hat{\beta}_0 + \hat{\beta}_1 X_o$$

όπου $X = 250$, ενώ το αντίστοιχο δ.ε. θα είναι (για μέση πρόβλεψη)

$$\hat{Y}_o \pm s(\hat{Y}_o) \cdot t_{n-2}(a/2) = \hat{\beta}_0 + \hat{\beta}_1 X_o \pm s \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{XX}}} \cdot t_{n-2}(a/2)$$

Για να βρούμε τα παραπάνω στο SPSS, εισάγουμε μία 16^η παρατήρηση στον SPSS data editor πληκτρολογώντας στην 16 γραμμή της στήλης X1 το 250 (οι υπόλοιπες θέσεις στην 16^η γραμμή αφήνονται κενές). Στη συνέχεια εκτελούμε τη διαδικασία της παλινδρόμησης **/Analyze/Regression/Linear** επιλέγοντας στο save τα **Unstandardized predicted values** ($\hat{Y}_o = \hat{\beta}_0 + \hat{\beta}_1 X_o$), **Prediction Intervals**. Στην 16^η στήλη λαμβάνονται τα αποτελέσματα:

Αναμενόμενες πωλήσεις: 154.9707

Δ.ε. 95% για την μέση πρόβλεψη: (151.5549, 158.3865)

Δ.ε. 95% για την μέση πρόβλεψη: (141.3418, 168,5996)

Αν σε μία πόλη πουλήθηκαν 120 μονάδες, ο πληθυσμός της εκτιμάται από το

$$X = \frac{Y - \hat{\beta}_0}{\hat{\beta}_1} = \frac{120 - 22,792}{0,52871} \approx 183.85$$

ε. Εφόσον το δ.ε. για τη μέση πρόβλεψη έχει τη μορφή

$$\hat{Y}_o \pm s(\hat{Y}_o) \cdot t_{n-2}(a/2) = \hat{\beta}_0 + \hat{\beta}_1 X_o \pm s \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{XX}}} \cdot t_{n-2}(a/2),$$

προφανώς θα έχει το μικρότερο εύρος όταν $X_o = \bar{X} = 241,733$ (από Descriptive Statistics). Για να βρούμε το δ.ε. σε αυτή την περίπτωση ακολουθούμε τα παραπάνω βήματα (για το δ.ε. όταν $X = 250$) αυτή τη φορά προσθέτοντας στην στήλη X1 την παρατήρηση 241,733.

Στ. Ο πίνακας ANOVA δίνεται απευθείας κατά την επιλογή Regression από το SPSS:

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	53416,719	1	53416,719	1432,139	,000 ^a
	Residual	484,881	13	37,299		
	Total	53901,600	14			

a. Predictors: (Constant), X1

b. Dependent Variable: Y

Ο παραπάνω πίνακας περιέχει τις ποσότητες:

<i>Model</i>	<i>SS</i>	<i>Df</i>	<i>MS</i>	<i>F</i>	<i>Sig.</i>
Regression	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$MSR = SSR$	$F^* = \frac{MSR}{MSE}$	$P(F > F^* F \sim F_{1,n-2})$
Residuals	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$		
Total	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n-1$			

Ο έλεγχος του μοντέλου μέσω του *F-test* μας οδηγεί στην απόρριψη της $H_0: \beta_1=0$ διότι το αντίστοιχο *p-value* είναι ίσο με 0,000.