

Π.Μ.Σ. ΒΙΟΣΤΑΤΙΣΤΙΚΗ
ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΔΙΑΣΠΟΡΑΣ
ΔΙΑΓΩΝΙΣΜΑ 10/2/2025

Τα αποτελέσματα των αναλύσεων δίνονται στο Παράρτημα

Πρόβλημα 1.

Σε μια στατιστική μελέτη αναλύθηκε ένα μοντέλο πολλαπλής παλινδρόμησης με εξαρτημένη μεταβλητή Y και ανεξάρτητες X_1, X_2 . Το μέγεθος δείγματος ήταν 40. Για το μοντέλο αυτό βρέθηκαν τα εξής:

$$\begin{aligned} R\text{-Square} &= 0.815, \\ SSY &= 5911.4. \end{aligned}$$

1. Να καταστρωθεί ο πίνακας ανάλυσης διασποράς.
2. Να γίνει έλεγχος σημαντικότητας για το συνολικό μοντέλο στο επίπεδο σημαντικότητας 5%.

Δίνονται οι παρακάτω κρίσιμες τιμές της κατανομής F

$$\begin{aligned} F_{0.05,3,38} &= 2.852, F_{0.05,38,3} = 8.598, & F_{0.05,2,39} &= 3.238, F_{0.05,39,2} = 19.470, \\ F_{0.05,2,37} &= 3.252, F_{0.05,37,2} &= 19.469 \end{aligned}$$

Πρόβλημα 2.

Σε μια μελέτη συγκεντρώθηκαν δεδομένα σχετικά με το μέγεθος του πληθυσμού (σε ζεύγη πτηνών) ενός είδους πτηνών (μεταβλητή pop) και την έκταση της αποικίας (σε km^2 , μεταβλητή area), για 22 αποικίες αυτού του πτηνού. Για την ανάλυση εκτελέστηκε το μοντέλο

$$\text{pop} = \beta_0 + \beta_1 \text{area} + \varepsilon$$

Στο παράρτημα δίνονται τα αποτελέσματα της παλινδρόμησης για το παραπάνω μοντέλο, το scatterplot pop-area με την αντίστοιχη γραμμή παλινδρόμησης, όπως επίσης και οι προβλεπόμενες τιμές ($\hat{y}$), τα τυπικά σφάλματα εκτίμησης και πρόβλεψης (stdp , stdf), τα κατάλοιπα jackknife και η απόσταση Cook για όλες τις παρατηρήσεις.

1. Να ερμηνεύσετε τις εκτιμήσεις ελαχίστων τετραγώνων.
2. Να εξετάσετε (σε επίπεδο σημαντικότητας 5%) ποιες παρατηρήσεις είναι outliers και ποιες influential (αν υπάρχουν).
3. Να υπολογίσετε και να ερμηνεύσετε (με όρους της εφαρμογής) ένα διάστημα πρόβλεψης για τη μέση τιμή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή πάρει την τιμή 1570.
4. Σε κάποιες παρατηρήσεις προκύπτουν αρνητικές προβλεπόμενες τιμές. Πού νομίζετε ότι οφείλεται αυτό; Πώς θα ελέγχατε αν η υπόθεσή σας ευσταθεί;

(Δίνεται $t_{0.025,20} = 2.086$).

Πρόβλημα 3.

Σε μια μελέτη συγκεντρώθηκαν δεδομένα σχετικά με τη θνησιμότητα και δημογραφικούς/οικονομικούς/περιβαλλοντικούς παράγοντες που ενδεχομένως να την επηρεάζουν.

pop_hse	Αριθμός ατόμων ανά οικογένεια
educ	Μέση διάρκεια μόρφωσης ατόμων άνω των 25
gdhse	Ποσοστό οικογενειών που ζουν σε αξιοπρεπή καταλύματα
nonw	Ποσοστό πληθυσμού που ανήκει σε μειονότητες
wcol	Ποσοστό εργαζομένων σε δουλειές γραφείου
poor	Ποσοστό οικογενειών με ετήσιο εισόδημα κάτω των \$3000
mort	Θνησιμότητα ανά 100.000 κατοίκους
popden	Πυκνότητα πληθυσμού (κάτοικοι ανά τετραγωνικό μίλι)
hc	Σχετικό επίπεδο μόλυνσης από υδρογονάνθρακες
nox	Σχετικό επίπεδο μόλυνσης από οξείδια του αζώτου
so2	Σχετικό επίπεδο μόλυνσης από διοξείδιο του θείου
smsa	Πόλη

1. Το Μοντέλο 1 είναι μοντέλο πολλαπλής παλινδρόμησης με εξαρτημένη μεταβλητή τη `mort` και ανεξάρτητες όλες τις υπόλοιπες με τη σειρά που δίνονται παραπάνω, εκτός από τη `smsa`. Για το μοντέλο αυτό δίνονται οι πίνακες F-tests type I (sequential) και F-tests type III (partial). Να δώσετε τις τιμές των δύο F-test για τη μεταβλητή `educ` και να εξηγήσετε πού οφείλεται η διαφορά τους.
2. Να υπολογίσετε το R^2 και τη στατιστική σημαντικότητα του μονοπαραγοντικού μοντέλου με εξαρτημένη μεταβλητή τη `mort` και ανεξάρτητη την `pop_hse` (δεν δίνεται πίνακας με αποτελέσματα αυτού του μοντέλου).
3. Το Μοντέλο 2 περιέχει τις `nonw`, `educ`, `poor` μόνο. Ζητείται να εκτιμήσετε κατά πόσο το Μοντέλο 2 είναι ισοδύναμο ή υποδεέστερο του Μοντέλου 1 ως προς τη στατιστική σημαντικότητα. Χρησιμοποιήστε μια μόνο (όποια θέλετε) μέθοδο για να απαντήσετε την ερώτηση.

Πρόβλημα 4.

Σε μια στατιστική ανάλυση εξετάστηκε η συσχέτιση μιας μεταβλητής Y με πιθανούς παράγοντες το φύλο (`sex`, 1=male) και την ηλικία (`ages`, κεντραρισμένη ως προς τη μέση τιμή =38.5).

Το μοντέλο κύριων επιδράσεων (Μοντέλο 1) και το μοντέλο αλληλεπιδράσεων (Μοντέλο 2) έδωσαν τα αποτελέσματα που φαίνονται στους αντίστοιχους πίνακες.

1. Για κάθε μοντέλο να γράψετε τις εξισώσεις παλινδρόμησης για κάθε φύλο. Να ερμηνεύσετε τις εξισώσεις παλινδρόμησης για κάθε φύλο για το Μοντέλο 2.
2. Στο Μοντέλο 1 οι κύριες επιδράσεις είναι μη στατιστικά σημαντικές και προκύπτει ότι κανένας από τους δύο παράγοντες δεν συσχετίζεται με την Y . Όμως στο Μοντέλο 2 και οι δυο κύριες επιδράσεις είναι στατιστικά σημαντικές. Πώς εξηγείτε αυτή τη διαφορά;

Παράρτημα: Αποτελέσματα Αναλύσεων

Πρόβλημα 2

```
. sum area pop
```

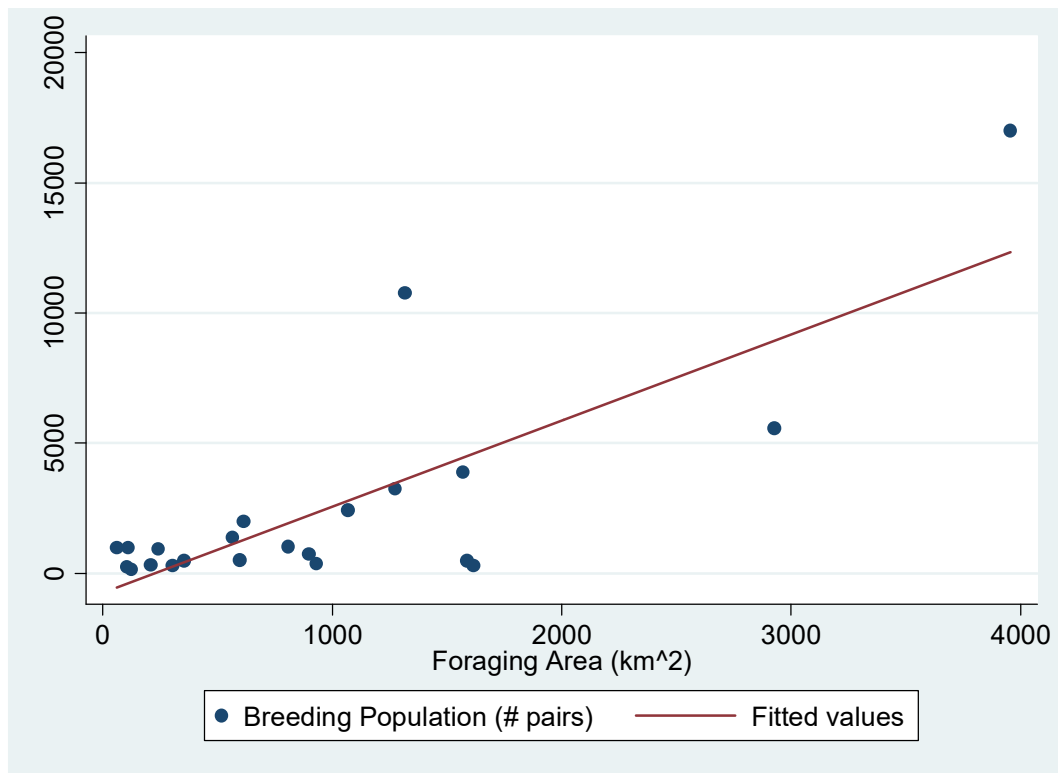
Variable	Obs	Mean	Std. Dev.	Min	Max
area	22	965.3418	965.0855	60.12	3957
pop	22	2452.818	4060.955	134	17000

```
. reg pop area
```

Source	SS	df	MS	Number of obs =	22
Model	213266632	1	213266632	F(1, 20) =	32.06
Residual	133051869	20	6652593.44	Prob > F =	0.0000
Total	346318501	21	16491357.2	R-squared =	0.6158
				Adj R-squared =	0.5966
				Root MSE =	2579.3

pop	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
area	3.302068	.5832034	5.66	0.000	2.085527 4.518609
_cons	-734.8063	786.9873	-0.93	0.362	-2376.433 906.8206

```
. tw (sc pop area) (lfit pop area)
```



. li area pop yhat stdp stdf jack cooksd

	area	pop	yhat	stdp	stdf	jackknife	cooksd
1.	207.7	311	-48.96672	705.4291	2673.99	.1414947	.000851
2.	1570	3872	4449.441	653.257	2660.703	-.2258667	.0018355
3.	1588	495	4508.878	658.983	2662.114	-1.681574	.0904687
4.	125.7	134	-319.7363	736.3281	2682.307	.179059	.0014948
5.	353.4	485	432.1446	655.5598	2661.269	.020652	.0000155
6.	931	372	2339.419	550.2651	2637.306	-.7728586	.0145341
7.	1616	284	4601.336	668.1206	2664.391	-1.832305	.1080087
8.	1317	10767	3614.017	586.9001	2645.193	3.600312	.2214471
9.	614	1975	1292.664	586.8357	2645.178	.2652833	.0020146
10.	60.12	970	-536.2859	762.2985	2689.552	.6014744	.0178831
11.	1273	3243	3468.727	578.433	2643.327	-.0875471	.0002135
12.	595.7	500	1232.236	590.6469	2646.027	-.2848652	.0023536
13.	105.6	250	-386.1079	744.1754	2684.472	.2514734	.0030123
14.	241.9	925	63.964	693.1102	2670.767	.3388219	.0046745
15.	111.1	970	-367.9465	742.0179	2683.875	.5318303	.0132352
16.	302.4	278	263.7391	672.2152	2665.421	.005582	1.20e-06
17.	808.9	1036	1936.237	557.418	2638.808	-.3495437	.0031305
18.	2927	5570	8930.347	1269.343	2874.687	-1.547974	.3579319
19.	1069	2430	2795.104	553.2136	2637.923	-.1413313	.0005064
20.	898.2	731	2231.111	551.293	2637.521	-.5855001	.0084843
21.	564.8	1364	1130.202	597.4599	2647.556	.0908401	.0002461
22.	3957	17000	12331.48	1829.351	3162.138	3.05651	3.336568

Πρόβλημα 3

Μοντέλο 1

. anova mort c.pop_hse c.educ c.gdhse c.nonw c.wcol c.poor c.popden c.hc c.nox c.so2

. anova, sequential

Number of obs = 60 R-squared = 0.7123
Root MSE = 36.6127 Adj R-squared = 0.6536

Source	Seq. SS	df	MS	F	Prob > F
Model	162613.088	10	16261.3088	12.13	0.0000
pop_hse	29110.7805	1	29110.7805	21.72	0.0000
educ	36980.8937	1	36980.8937	27.59	0.0000
gdhse	5157.26091	1	5157.26091	3.85	0.0555
nonw	57494.4153	1	57494.4153	42.89	0.0000
wcol	106.266966	1	106.266966	0.08	0.7795
poor	16928.1928	1	16928.1928	12.63	0.0009
popden	2386.95033	1	2386.95033	1.78	0.1882
hc	1511.95497	1	1511.95497	1.13	0.2934
nox	12712.3825	1	12712.3825	9.48	0.0034
so2	223.989476	1	223.989476	0.17	0.6845
Residual	65683.8336	49	1340.4864		
Total	228296.921	59	3869.43934		

. anova, partial

Number of obs = 60 R-squared = 0.7123
Root MSE = 36.6127 Adj R-squared = 0.6536

Source	Partial SS	df	MS	F	Prob > F
Model	162613.088	10	16261.3088	12.13	0.0000
pop_hse	2032.01806	1	2032.01806	1.52	0.2241
educ	3798.45668	1	3798.45668	2.83	0.0987
gdhse	2473.37798	1	2473.37798	1.85	0.1806
nonw	37814.1504	1	37814.1504	28.21	0.0000
wcol	633.507323	1	633.507323	0.47	0.4950
poor	6530.1437	1	6530.1437	4.87	0.0320
popden	901.357178	1	901.357178	0.67	0.4162
hc	5073.07303	1	5073.07303	3.78	0.0575
nox	3950.72651	1	3950.72651	2.95	0.0923
so2	223.989476	1	223.989476	0.17	0.6845
Residual	65683.8336	49	1340.4864		
Total	228296.921	59	3869.43934		

Μοντέλο 3

. reg mort nonw educ poor

Source	SS	df	MS	Number of obs =	60
Model	139242.252	3	46414.0839	F(3, 56) =	29.19
Residual	89054.6696	56	1590.26196	Prob > F =	0.0000
				R-squared =	0.6099
				Adj R-squared =	0.5890
Total	228296.921	59	3869.43934	Root MSE =	39.878

mort	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nonw	5.415529	.8260849	6.56	0.000	3.760682	7.070376
educ	-35.52083	6.757944	-5.26	0.000	-49.05862	-21.98304
poor	-4.968832	1.893301	-2.62	0.011	-8.76157	-1.176094
_cons	1337.283	85.51945	15.64	0.000	1165.967	1508.599

Πρόβλημα 4

MONTEΛΟ 1

.reg y agec sex

Source	SS	df	MS	Number of obs = 48		
Model	621.145253	2	310.572626	F(2, 45)	=	1.61
Residual	8699.32966	45	193.318437	Prob > F	=	0.2119
Total	9320.47491	47	198.307977	R-squared	=	0.0666
				Adj R-squared	=	0.0252
				Root MSE	=	13.904

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
agec	-.2708372	.1759272	-1.54	0.131	-.6251726	.0834983
sex	4.190217	4.11291	1.02	0.314	-4.093609	12.47404
_cons	57.45027	3.194067	17.99	0.000	51.01709	63.88345

MONTEΛΟ 2

.gen sagc=sex*agec

.reg y agec sex sagc

Source	SS	df	MS	Number of obs = 48		
Model	8124.75017	3	2708.25006	F(3, 44)	=	99.66
Residual	1195.72474	44	27.1755623	Prob > F	=	0.0000
Total	9320.47491	47	198.307977	R-squared	=	0.8717
				Adj R-squared	=	0.8630
				Root MSE	=	5.213

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
agec	1.17944	.1093996	10.78	0.000	.9589592	1.39992
sex	3.33729	1.542916	2.16	0.036	.2277468	6.446833
sagc	-2.278622	.1371282	-16.62	0.000	-2.554986	-2.002259
_cons	58.81506	1.200371	49.00	0.000	56.39587	61.23425