

Ανάλυση Δεδομένων με Η/Υ

ΘΕΜΑ 1 (1.5 μονάδες): Γράψτε μια συνάρτηση η οποία θα δέχεται ως όρισμα ένα διάνυσμα x μήκους n με στοιχεία x_1, \dots, x_n και έναν θετικό ακέραιο αριθμό $k \leq n$ και θα επιστρέφει υπό μορφή διανύσματος τις τιμές του κινητού μέσου k όρων, δηλαδή το διάνυσμα με στοιχεία

$$\frac{x_1 + x_2 + \dots + x_k}{k}, \frac{x_2 + x_3 + \dots + x_{k+1}}{k}, \dots, \frac{x_{n-k+1} + x_{n-k+2} + \dots + x_n}{k}.$$

Αν το $k > n$ ή αν το k δεν είναι θετικός ακέραιος αριθμός, η συνάρτησή σας θα επιστρέψει ένα κατάλληλο μήνυμα λάθους και δε θα υπολογίζει τίποτα.

ΘΕΜΑ 2 (5.5 μονάδες): Μια μελέτη του Τμήματος Ιστορίας της Τέχνης στοχεύει στην αξιολόγηση τεσσάρων χαρακτηριστικών των έργων 54 κλασικών ζωγράφων (σε κλίμακα από το 0 ως το 20). Συλλέχτηκαν δεδομένα για τις ακόλουθες μεταβλητές: **Name** = όνομα του ζωγράφου, **Composition** = βαθμολογία της σύνθεσης, **Drawing** = βαθμολογία του σχεδίου, **Colour** = βαθμολογία του χρώματος, **Expression** = βαθμολογία της έκφρασης και **School** = η καλλιτεχνική σχολή στην οποία ανήκει κάθε καλλιτέχνης (A : Αναγέννηση, B : Μανιερισμός, C : Μπαρόκ, D : Ενετική σχολή, E : Λούβαρδική σχολή, F : 16^{ος} αιώνας, G : 17^{ος} αιώνας, H : Γαλλική σχολή).

(Α) Με ποιους τρόπους (αριθμητικούς και γραφικούς) και με ποιες εντολές στην R θα περιγράφατε τις τιμές κάθε μεταβλητής στο δείγμα;

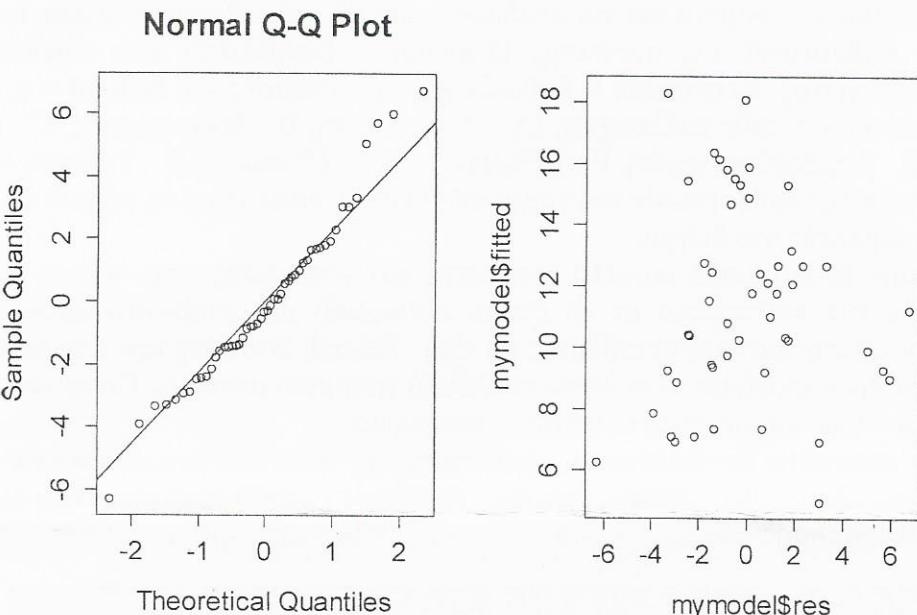
(Β) Προσαρμόσαμε το πολλαπλό μοντέλο γραμμικής παλινδρόμησης στην R (και το αποτέλεσμα καταχωρήθηκε σε ένα αντικείμενο με το όνομα **mymodel**) με μεταβλητή απόκρισης την τ.μ. **Composition** και επεξηγηματικές μεταβλητές τις εξής: **School**, **Drawing** και **Expression**. Με ποια εντολή στην R θα προσαρμόζατε το εν λόγω πολλαπλό γραμμικό μοντέλο; Ποιες είναι οι εικονικές μεταβλητές στο μοντέλο και με ποιες κατηγορίες αναφοράς;

(Γ) Με βάση τα παρακάτω αποτελέσματα, ερμηνεύστε τις τιμές των εκτιμητών των συντελεστών του γραμμικού μοντέλου καθώς και την τιμή του συντελεστή προσδιορισμού. Ποιο είναι το τυπικό σφάλμα της παλινδρόμησης; *and also*

Call:	lm(formula = Composition ~ Expression + Drawing + School)																																																							
Residuals:																																																								
<code>Min</code> <code>1Q</code> <code>Median</code> <code>3Q</code> <code>Max</code>																																																								
-6.2811 -1.5203 -0.2999 1.5680 6.7493																																																								
Coefficients:																																																								
<table> <thead> <tr> <th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr> </thead> <tbody> <tr> <td>(Intercept)</td><td>4.6243</td><td>2.4138</td><td>1.916</td><td>0.062564 .</td></tr> <tr> <td>Expression</td><td>0.4840</td><td>0.1304</td><td>3.712</td><td>0.000626 ***</td></tr> <tr> <td>Drawing</td><td>0.1105</td><td>0.1844</td><td>0.599</td><td>0.552640</td></tr> <tr> <td>SchoolB</td><td>1.6100</td><td>1.6297</td><td>0.988</td><td>0.329144</td></tr> <tr> <td>SchoolC</td><td>3.5826</td><td>1.5482</td><td>2.314</td><td>0.025885 *</td></tr> <tr> <td>SchoolD</td><td>1.8334</td><td>1.5006</td><td>1.222</td><td>0.228945</td></tr> <tr> <td>SchoolE</td><td>3.5859</td><td>1.5165</td><td>2.365</td><td>0.022987 *</td></tr> <tr> <td>SchoolF</td><td>-2.2574</td><td>1.9328</td><td>-1.168</td><td>0.249744</td></tr> <tr> <td>SchoolG</td><td>3.8219</td><td>1.9752</td><td>1.935</td><td>0.060084 .</td></tr> <tr> <td>SchoolH</td><td>1.7794</td><td>1.8942</td><td>0.939</td><td>0.353161</td></tr> </tbody> </table>			Estimate	Std. Error	t value	Pr(> t)	(Intercept)	4.6243	2.4138	1.916	0.062564 .	Expression	0.4840	0.1304	3.712	0.000626 ***	Drawing	0.1105	0.1844	0.599	0.552640	SchoolB	1.6100	1.6297	0.988	0.329144	SchoolC	3.5826	1.5482	2.314	0.025885 *	SchoolD	1.8334	1.5006	1.222	0.228945	SchoolE	3.5859	1.5165	2.365	0.022987 *	SchoolF	-2.2574	1.9328	-1.168	0.249744	SchoolG	3.8219	1.9752	1.935	0.060084 .	SchoolH	1.7794	1.8942	0.939	0.353161
	Estimate	Std. Error	t value	Pr(> t)																																																				
(Intercept)	4.6243	2.4138	1.916	0.062564 .																																																				
Expression	0.4840	0.1304	3.712	0.000626 ***																																																				
Drawing	0.1105	0.1844	0.599	0.552640																																																				
SchoolB	1.6100	1.6297	0.988	0.329144																																																				
SchoolC	3.5826	1.5482	2.314	0.025885 *																																																				
SchoolD	1.8334	1.5006	1.222	0.228945																																																				
SchoolE	3.5859	1.5165	2.365	0.022987 *																																																				
SchoolF	-2.2574	1.9328	-1.168	0.249744																																																				
SchoolG	3.8219	1.9752	1.935	0.060084 .																																																				
SchoolH	1.7794	1.8942	0.939	0.353161																																																				
<code>---</code>																																																								
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1																																																								
Residual standard error: 2.92 on 40 degrees of freedom																																																								
Multiple R-squared: 0.6016, Adjusted R-squared: 0.512																																																								
F-statistic: 6.711 on 9 and 40 DF, p-value: 8.537e-06																																																								

- (Δ) Δώστε τον τύπο του συντελεστή προσδιορισμού καθώς και τον τύπο του διορθωμένου συντελεστή προσδιορισμού στο παραπάνω μοντέλο. Ποια είναι η χρησιμότητα του τελευταίου;
- (Ε) Ποιον έλεγχο εξετάζουμε με το F-test στην παραπάνω πολλαπλή γραμμική παλινδρόμηση; Με βάση τα παραπάνω αποτελέσματα σε τι τελικά συμπεράσματα θα καταλήγατε;
- (ΣΤ) Εκτιμήστε την αναμενόμενη τιμή της βαθμολογίας για τη σύνθεση (μεταβλητή **Composition**) ενός ζωγράφου της Λομβαρδικής σχολής με βαθμολογία σχεδίου 9 και βαθμολογία έκφρασης 16. Δώστε και την αντίστοιχη εντολή στην R.

Τα κάτωθι διαγράμματα απεικονίζουν τη γραφική παράσταση των δειγματικών ποσοστημορίων της κατανομής των υπολοίπων του μοντέλου ως προς τα θεωρητικά ποσοστημόρια της Κανονικής κατανομής (αριστερά) και τις προσαρμοσμένες τιμές του μοντέλου έναντι των υπολοίπων (δεξιά). Δώστε τις εντολές της R που δημιουργούν τα εν λόγω διαγράμματα. Αναφέρατε όλες τις προϋποθέσεις που πρέπει να ισχύουν για το παραπάνω μοντέλο, και σχολιάστε (χρησιμοποιώντας την πληροφορία που σας δίνεται από τα παρακάτω διαγράμματα και την περιγραφή του προβλήματος) αν το μοντέλο που προσαρμόσαμε είναι κατάλληλο για τον σκοπό της μελέτης.



- (Η) Με ποιες εντολές θα μπορούσατε να ελέγξετε αν προκύπτει πρόβλημα πολυσυγγραμμικότητας στο πολλαπλό γραμμικό μοντέλο που προσαρμόσατε προηγουμένως; Ποια θα ήταν μια πιθανή λύση σε ενδεχόμενο τέτοιο πρόβλημα;

ΘΕΜΑ 3 (3 μονάδες):

Στα πλαίσια του υπολογισμού του γενικού δείκτη τιμών υλικών κατασκευής νέων κτιρίων (π.χ. σκυρόδεμα, ξυλεία οικοδομών, γυψοσανίδες, παράθυρα, παρκέτα, ενεργειακοί υαλοπίνακες, κλπ), καταγράφηκε το μέσο ποσοστό μεταβολής των τιμών υλικών κατασκευής κατά τα έτη 2013 και 2014 σε σχέση με το 2012 για 12 κατασκευαστές-εργολήπτες τυχαία επιλεγμένους από διάφορες περιοχές της Ελλάδας.

2013	0.7	-2.1	-3.2	0.2	-0.8	-0.4	1.1	-0.3	-2.6	-0.9	0.7	-1.3
2014	-1.2	-3.6	-1.5	-5.9	-1.0	0.1	-4.1	-3.4	-7.1	-1.5	0.3	-2.2

Θέλουμε να ελέγξουμε, σε ε.σ. 2%, την υπόθεση ότι η μέση τιμή του κόστους υλικών κατασκευής νέων κτιρίων σε σχέση με το 2012 δε διαφέρει σημαντικά ανάμεσα στα έτη 2013 και 2014, με εναλλακτική ότι είναι χαμηλότερη για το 2014.

- i) Τι είδους στατιστική ανάλυση θα εφαρμόζατε και με ποιες εντολές στην R;
- ii) Ποιες προϋποθέσεις θα ελέγχατε για τη στατιστική ανάλυση του παραπάνω ερωτήματος και με ποιες εντολές στην R;
- iii) Αν δεν ίσχυαν οι παραπάνω προϋποθέσεις, ποιον έλεγχο θα εφαρμόζατε και με ποια εντολή στην R;

Διάρκεια Εξέτασης: 2 ½ ώρες

EYXOMAI EΠΙΤΥΧΙΑ