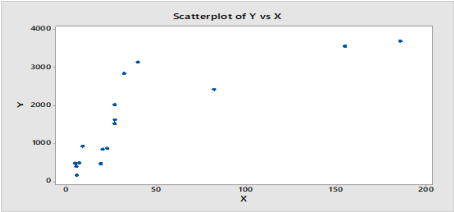
ΕΞΕΤΑΣΗ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ, ΣΑΧΜ,
18/06/21

1. (**20 μονάδες**) Απαντήσετε στα παρακάτω. Κάθε ερώτημα περιέχει συνολικά 5 μονάδες.

1. Έστω το παρακάτω διάγραμμα διασποράς μεταξύ των τιμών της ανεξάρτητης μεταβλητής X και της εξαρτημένης μεταβλητής Y.



Ποιος θεωρείτε ότι θα ήταν ο καταλληλότερος μετασχηματισμός στην X προκειμένου να μπορέσουμε να προσαρμόσουμε μοντέλο απλής γραμμικής παλινδρόμησης και γιατί;

2. Όταν τα σφάλματα ακολουθούν την κανονική κατανομή, τι κατανομή ακολουθούν προσεγγιστικά τα Studentized υπόλοιπα και τι όταν το μέγεθος του δείγματος τείνει στο άπειρο;

3. Σε καθεμία από τις παρακάτω περιπτώσεις σημειώστε το γράμμα Σ αν ο ισχυρισμός είναι σωστός ή το γράμμα Λ, αν ο ισχυρισμός είναι λανθασμένος.

   (a) Ο έλεγχος Kolmogorov-Smirnov χρησιμοποιείται για να ελέγξουμε την ανεξαρτησία των σφαλμάτων του απλού γραμμικού μοντέλου.

   (b) Τόσο τα σφάλματα (ε_i) όσο και τα υπόλοιπα ($\hat{\varepsilon}_i$) έχουν σταθερή διακύμανση.

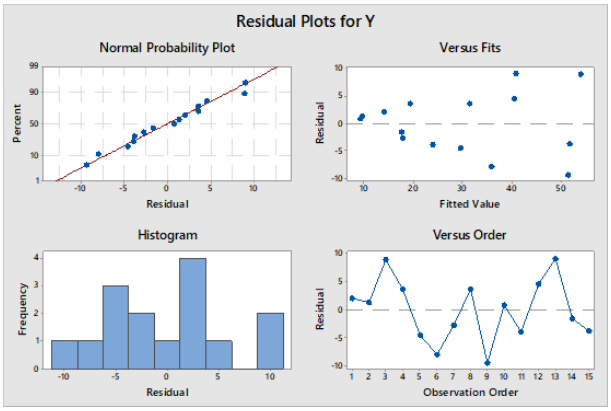
   (c)Από την ανάλυση υπολοίπων ενός απλού γραμμικού μοντέλου το οποίο προσαρμόσαμε στα δεδομένα μας, πήραμε το παρακάτω output.

Mean	0,1509
StDev	0,8482
N	16
Anderson-Darling	0,223
P-Value	0,792

   Σε επίπεδο σημαντικότητας 10% δεν απορρίπτουμε την υπόθεση της κανονικότητας των σφαλμάτων.

   (d)Αν από το διάγραμμα διασποράς n παρατηρήσεων διαπιστώσουμε ότι η μορφή της συναρτησιακής σχέσης μεταξύ των X και Y περιγράφεται ικανοποιητικά από τη σχέση $y = \gamma_0\gamma_1^x$, τότε θα πρέπει να μετασχηματίσουμε τα δεδομένα που αφορούν την μεταβλητή απόκρισης Y σύμφωνα με τον τύπο $Y' = 1/Y$.

4. Από την ανάλυση υπολοίπων ενός απλού γραμμικού μοντέλου το οποίο προσαρμόσαμε στα δεδομένα μας, πήραμε τα παρακάτω διαγράμματα. Σχολιάστε αν πληρούνται οι προϋποθέσεις της κανονικότητας, της ομοσκεδαστικότητας και της ανεξαρτησίας των σφαλμάτων.



2. (**40 μονάδες**) Θεωρήστε το απλό γραμμικό μοντέλο :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,\ i = 1, \ldots, 100,$$

με ομοσκεδαστικά και ασυσχέτιστα σφάλματα που ακολουθούν κανονική κατανομή.

a. (2 μονάδες) Γράψτε την κατανομή του $\hat{\boldsymbol{\beta}}$, του εκτιμητή ελαχίστων τετραγώνων του διανύσματος των συντελεστών, $\boldsymbol{\beta}$.

b. (3 μονάδες) Γράψτε την κατανομή της s^2, της αμερόληπτης εκτιμήτριας της διακύμανσης των σφαλμάτων.

c. (15 μονάδες) Αποδείξτε πως $\hat{\beta}_1 - \hat{\beta}_3$ και s^2 είναι στατιστικά ανεξάρτητες τυχαίες μεταβλητές. Δείξετε <u>ΟΛΑ</u> τα βήματα.

d. (20 μονάδες) Έστω $W_i = \frac{\hat{Y}_i}{s}$. Βρείτε τα παρακάτω:

$$E(\sum_{i=1}^{100} W_i^2) \quad , \quad Var(\sum_{i=1}^{100} W_i^2)$$

3. (**40 μονάδες**) Χρησιμοποιήστε το output που αναφέρεται σε αυτή την άσκηση (στο τέλος της εξέτασης). Η απόκριση είναι ο λογάριθμος της τιμής της σπιρομέτρησης και οι συμμεταβλητή είναι το ύψος (ht, σε ίντσες) , η ηλικία (age,σε έτη) και το φύλο (sex, 0 για τα κορίτσια κι 1 για τα αγόρια) . Τα δεδομένα αναφέρονται σε μη-καπνιστές (αγόρια και κορίτσια). Υποθέσετε πως τα σφάλματα είναι τυχαίο δειγμα από κανονική κατανομή. Στο πλήρες μοντέλο έχουν συμπεριληφθεί και αλληλεπιδράσεις του φύλου με την ηλικία και το ύψος.

a. (5 μονάδες) Χρησιμοποιώντας το μερικό (partial) F έλεγχο, ελέγξτε την μηδενική υπόθεση που λέει πως το περιορισμένο μοντέλο με μοναδική συμμεταβλητή την ηλικία είναι επαρκές (σε σχέση με το πλήρες μοντέλο). Γράψτε αναλυτικά τις υποθέσεις, δώστε τη τιμή της σ.σ.ε., προσδιορίσετε το $p-value$, αναφέρετε την απόφαση και γράψτε το συμπέρασμα σε απλά ελληνικά.

b. (5 μονάδες) Ελέγξτε την υπόθεση πως, βάσει του πλήρους μοντέλου, το φύλο δεν επιρεάζει την επίδραση της ηλικίας στην απόκριση. Γράψτε αναλυτικά τις υποθέσεις, δώστε τη τιμή της

σ.σ.ε., προσδιορίσετε το $p - value$, αναφέρετε την απόφαση και γράψτε το συμπέρασμα σε απλά ελληνικά.

c. (10 μονάδες) Υπολογίστε ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή της απόκρισης του πληθυσμού 14-χρονων κοριτσιών με ύψος 62 ίντσες.

d. (10 μονάδες) Υπολογίστε ένα 95% διάστημα πρόβλεψης για τη τιμή της απόκρισης ενός τυχαία επιλεγμένου 14-χρονου κοριτσιού με ύψος 62 ίντσες.

e. (10 μονάδες) Βάσει του πλήρους μοντέλου **Γράψτε** το (προσεγγιστικό) 95% Δ.Ε. για την μέση διαφορά στη τιμή της σπιρομέτρησης μεταξύ 14-χρονων αγοριών με ύψος 64 ίντσες και 15-χρονων αγοριών με ύψος 66 ίντσες.

Οι πράξεις να γίνονται με ακρίβεια 6 δεκαδικών.

Output για 3η άσκηση

NON-SMOKERS

Regression Analysis: ln(FEV) versus age; ht; sex; age*sex; ht*sex

Regression Equation

ln(FEV) = -1,888 + 0,02708 age + 0,04129 ht - 0,007 sex - 0,00263 age*sex + 0,00101 ht*sex

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-1,888	0,138	(-2,158; -1,617)	-13,71	0,000	
age	0,02708	0,00504	(0,01717; 0,03698)	5,37	0,000	5,51
ht	0,04129	0,00287	(0,03565; 0,04692)	14,40	0,000	7,63
sex	-0,007	0,175	(-0,350; 0,337)	-0,04	0,969	220,33
age*sex	-0,00263	0,00741	(-0,01718; 0,01191)	-0,36	0,722	43,51
ht*sex	0,00101	0,00374	(-0,00634; 0,00835)	0,27	0,788	390,07

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0,142745	81,63%	81,48%	12,1552	81,21%	-613,54	-583,08

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Seq MS	F-Value	P-Value
Regression	5	52,8025	81,63%	52,8025	10,5605	518,28	0,000
age	1	40,6713	62,88%	0,5872	40,6713	1996,03	0,000
ht	1	12,0111	18,57%	4,2259	12,0111	589,47	0,000
sex	1	0,1175	0,18%	0,0000	0,1175	5,77	0,017
age*sex	1	0,0011	0,00%	0,0026	0,0011	0,05	0,815
ht*sex	1	0,0015	0,00%	0,0015	0,0015	0,07	0,788
Error	583	11,8793	18,37%	11,8793	0,0204		
Lack-of-Fit	307	6,3938	9,89%	6,3938	0,0208	1,05	0,346
Pure Error	276	5,4854	8,48%	5,4854	0,0199		
Total	588	64,6818	100,00%				

Tests use the sequential sums of squares

$(X^T X)^{-1} =$

0,930937	0,0211478	-0,0188811	-0,93094	-0,0211478	0,0188811
0,021148	0,0012485	-0,0005510	-0,02115	-0,0012485	0,0005510
-0,018881	-0,0005510	0,0004033	0,01888	0,0005510	-0,0004033
-0,930937	-0,0211478	0,0188811	1,50046	0,0403152	-0,0311044
-0,021148	-0,0012485	0,0005510	0,04032	0,0026919	-0,0010898
0,018881	0,0005510	-0,0004033	-0,03110	-0,0010898	0,0006869