

**ΕΞΕΤΑΣΗ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ, ΣΑΧΜ,
12/02/2021**

1. (20 μονάδες) Θεωρήστε το απλό γραμμικό μοντέλο :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n,$$

με ομοσχεδαστικά και ασυσχέτιστα σφάλματα.

- a. (5 μονάδες) Γράψτε **αναλυτικά** την LS εκτίμηση του $\theta = \beta_0 + 2\beta_1$ ως γραμμικό συνδυασμό των αποκρίσεων, **δηλαδή** στην μορφή $\hat{\theta}_{LS} = \sum w_i Y_i$.

- b. (3 μονάδες) Γράψτε **αναλυτικά** τα υπόλοιπα $\hat{\epsilon}_i$ ως γραμμικούς συνδυασμούς των αποκρίσεων, **δηλαδή** στην μορφή $\sum w_i Y_i$.

- c. (7 μονάδες) **Δείξτε** πως

$$\text{cov}\left(\sum_{i=1}^n \hat{\epsilon}_i, \hat{\theta}_{LS}\right) = \text{cov}\left(\sum_{i=1}^n x_i \hat{\epsilon}_i, \hat{\theta}_{LS}\right)$$

- d. (5 μονάδες) **Βρείτε** αμερόληπτη εκτιμήτρια της διασποράς των σφαλμάτων. Πρέπει να **αποδείξετε** πως η εκτιμήτριά σας είναι αμερόληπτη ΧΩΡΙΣ να υποθέσετε κανονικότητα.

2. (20 μονάδες) Θεωρήστε το κανονικό γραμμικό μοντέλο :

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \end{aligned} \quad (1)$$

όπου για τον $n \times p$ πίνακα συμμεταβλητών \mathbf{X} ισχύει $\text{rank}(\mathbf{X}) = p$.

- a. (10 μονάδες) **Δείξτε** πως ο εκτιμητής ελαχίστων τετραγώνων, $\hat{\boldsymbol{\beta}}$ και η 'κλασσική' αμερόληπτη εκτίμηση της διασποράς των σφαλμάτων, s^2 , είναι ανεξάρτητα.

- b. (10 μονάδες) **Βρείτε** (δείξτε όλα τα βήματα) την αναμενόμενη τιμή

$$E\left(\frac{\hat{\boldsymbol{\beta}}^T (\mathbf{X}^T \mathbf{X}) \hat{\boldsymbol{\beta}}}{\sigma^2}\right)$$

3. (20 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). Τα δεδομένα αναφέρονται σε 310 αγόρια μη-καπνιστές. Υποθέσετε πως τα σφάλματα είναι τυχαίο δείγμα από κανονική κατανομή.
- (10 μονάδες) **Γράψτε** εκτίμηση της επίδρασης του ύψους στη τιμή της απόκρισης. Ερμηνεύστε την.
 - (10 μονάδες) Απλά **Γράψτε** το 90% Διάστημα Πρόβλεψης για ένα τυχαία επιλεγμένο αγόρι (μη-καπνιστή) με $\text{ύψος}=61,519$ ίντσες. Ερμηνεύστε το.
4. (20 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 συμμεταβλητές είναι το ύψος (ht, σε ίντσες) και η ηλικία (σε έτη). Τα δεδομένα αναφέρονται σε μη-καπνιστές (αγόρια και κορίτσια). Υποθέσετε πως τα σφάλματα είναι τυχαίο δείγμα από κανονική κατανομή.
- (5 μονάδες) **Γράψτε** την τιμή της εκτίμησης της συνδιακύμανσης μεταξύ της εκτίμησης του συντελεστή της ηλικίας και της εκτίμησης του συντελεστή του ύψους.
 - (15 μονάδες) Απλά **Γράψτε** το 95% Διάστημα Πρόβλεψης για την απόκριση για νέο άτομο 15 ετών με $\text{ύψος}=64$ ίντσες.
5. (20 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 δείγμα από κανονική κατανομή. Στο πλήρες μοντέλο έχουν συμπεριληφθεί και αλληλεπιδράσεις του φύλου με την ηλικία και το ύψος.
- (10 μονάδες) **Γράψτε** μια εκτίμηση για την μέση διαφορά στη τιμή της σπιρομέτρησης μεταξύ 12-χρονων αγοριών με ύψος 59 ίντσες και 12-χρονων κοριτσιών με το ίδιο ύψος .
 - (10 μονάδες) Βάσει του πλήρους μοντέλου **Γράψτε** το (προσεγγιστικό) 98% Δ.Ε. για την μέση διαφορά στη τιμή της σπιρομέτρησης

σης μεταξύ 12-χρονων αγοριών με ύψος 59 ίντσες και 12-χρονων κοριτσιών με το ίδιο ύψος.

Οι πράξεις να γίνονται με ακρίβεια 6 δεκαδικών.

Οι απαντήσεις στις ασκήσεις 3,4,5 να δίνονται στο word (όχι φωτογραφίες).
Φωτογραφία συνεπάγεται μηδενισμός της άσκησης.

ΔΙΑΡΚΕΙΑ ΕΞΕΤΑΣΗΣ: 2 ΩΡΕΣ.

Output για 3^η Άσκηση

MALE NON-SMOKERS

Descriptive Statistics: ht; ln(FEV)

Statistics

Variable	N	Mean	Variance	Sum of Squares
ht	310	61,519	39,286	1185375,000
ln(FEV)	310	0,9439	0,1253	314,9182

R Large residual

MALE NON-SMOKERS

NON-SMOKERS

Descriptive Statistics: ht*ln(FEV)

Statistics

Variable	N	Mean	Variance	Sum of Squares
ht*ln(FEV)	310	60,08	742,31	1348478,41

Output για 4^η άσκηση

NON-SMOKERS

Regression Analysis: ln(FEV) versus age; ht

Regression Equation

$$\ln(\text{FEV}) = -1,9333 + 0,02469 \text{ age} + 0,04267 \text{ ht}$$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-1,9333	0,0816	?	?
age	0,02469	0,00365	?	?
ht	0,04267	0,00176	?	?

Model Summary

S	R-sq	R-sq(adj)
0,143097	?	81,39%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	52,68	26,3412	1286,40	0,000
Error	586	?	0,0205		
Total	588	64,68			

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

0,324951	0,0091545	-0,0067732
0,009155	0,0006493	-0,0002532
-0,006773	-0,0002532	0,0001516

Output για 5^η άσκηση

NON-SMOKERS

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

Regression Equation

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

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