ΕΞΕΤΑΣΗ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ, ΣΑΧΜ, 03/09/20

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

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

με ομοσκεδαστικά και ασυσχέτιστα σφάλματα.

   a. (2 μονάδες) Γράψτε **αναλυτικά** τις εκτιμήσεις των συντελεστών β_0, β_1 ως γραμμικούς συνδυασμούς των αποκρίσεων, **δηλαδή** στην μορφή $\sum w_i Y_i$.

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

   c. (5 μονάδες) **Βρείτε** το $cov(\hat{\epsilon}_i, \sum \hat{Y}_j)$

   d. (10 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 :

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

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

   a. (8 μονάδες) Για 2 τυχαία διανύσματα $\boldsymbol{U}, \boldsymbol{V}$ και 2 σταθερούς πίνακες $\boldsymbol{B}, \boldsymbol{C}$ ισχύει $Cov(\boldsymbol{BU}, \boldsymbol{CV}) = \boldsymbol{B}Cov(\boldsymbol{U}, \boldsymbol{V})\boldsymbol{C}^T$. **Χρησιμοποιήστε** αυτό το αποτέλεσμα για να **βρείτε** την συνδιακύμανση μεταξύ του εκτιμητή ελαχίστων τετραγώνων και των υπολοίπων.

   b. (5 μονάδες) **Δείξτε** πως ο εκτιμητής ελαχίστων τετραγώνων, $\hat{\boldsymbol{\beta}}$ και η 'κλασσική' αμερόληπτη εκτίμηση της διασποράς των σφαλμάτων, s^2, είναι ανεξάρτητα.

   c. (7 μονάδες) Υποθέτοντας πως $\boldsymbol{\beta} = \mathbf{0}$, **βρείτε** (**δείξτε** όλα τα βήματα, αφού πρώτα απλά **αναφέρετε** την κατανομή της s^2) την κατανομή της

$$\frac{\hat{\boldsymbol{\beta}}^T(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}}}{ps^2}$$

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 δειγμα από κανονική κατανομή.

   a. (4 μονάδες) Αφού υπολογίσετε τις τιμές των συντελεστών **Γράψτε** την προσαρμοσμένη ευθεία παλιδρόμησης.

   b. (2 μονάδες) **Γράψτε** την τιμή του SSR (άθροισμα τετραγώνων παλινδρόμησης).

   c. (2 μονάδες) **Γράψτε** την τιμή του SSE (άθροισμα τετραγώνων υπολοίπων).

   d. (6 μονάδες) Απλά **Γράψτε** το 99% Δ.Ε. για την μέση απόκριση για ύψος=61,519. Ερμηνεύστε το.

   e. (6 μονάδες) Απλά **Γράψτε** το 99% Διάστημα Πρόβλεψης για ένα τυχαία επιλεγμένο αγόρι (μη-καπνιστή) με 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 συμμεταβλητή 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 δειγμα από κανονική κατανομή.

   a. (5 μονάδες) **Γράψτε** την τιμή της εκτίμησης της συνδιακύμανσης μεταξύ της εκτίμησης του συντελεστή της ηλικίας και της εκτίμησης του συντελεστή του ύψους.
   Εξίσωση παλινδρόμησης: $E(lnFEV) = \beta_0 + \beta_1 age + \beta_2 ht$.
   Θέλω να υπολογίσω $\hat{cov}(\hat{\beta}_1, \hat{\beta}_2)$.

   $\hat{cov}(\hat{\beta}_1, \hat{\beta}_2) = 0.0205(-0.0002532)$.

   b. (15 μονάδες) Απλά **Γράψτε** το 99% Δ.Ε. για την μέση απόκριση για άτομα 15 ετών με ύψος=64 ίντσες.
   $\delta = \beta_0 + 15\beta_1 + 64\beta_2$.

$\hat{\delta} = \hat{\beta}_0 + 15\hat{\beta}_1 + 64\hat{\beta}_2 = -1.9333 + 15(0.02469) + 64(0.04267) = 1.16793$

$var(\hat{\delta}) = var(\hat{\beta}_0) + 225var(\hat{\beta}_1) + 4096var(\hat{\beta}_2) + 30cov(\hat{\beta}_0, \hat{\beta}_1) + 128cov(\hat{\beta}_0, \hat{\beta}_2) + 1920cov(\hat{\beta}_1, \hat{\beta}_2)$.

$\hat{var}(\hat{\delta}) = 0.0205 \times [0.324951 + 225(0.0006493) + 4096(0.0001516) + 30(0.0091545) - 128(0.0067732) - 1920(0.0002532)] = 0.0002771293$

99% ΔE: $1.16793 \pm 2.576(0.0166472) : [1.125047, 1.210813]$

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 δειγμα από κανονική κατανομή. Στο πλήρες μοντέλο έχουν συμπεριληφθεί και αλληλεπιδράσεις του φύλου με την ηλικία και το ύψος.

   a. (3 μονάδες) Θεωρήσετε τη μηδενική υπόθεση που λέει πως το φύλο δεν χρειάζεται να ληφθεί υπόψη όταν πρόκειται για πρόβλεψη της τιμής της σπιρομέτρησης ενός ατόμου έναντι της δίπλευρης εναλλακτικής. **Γράψτε** την τιμή της σ.σ.ε., προσέγγιση του p-value, συμπέρασμα σε απλά ελληνικά.

   Πλήρες μοντέλο: $E(lnFEV) = \beta_0 + \beta_1 age + \beta_2 ht + \beta_3 sex + \beta_4(age \times sex) + \beta_5(ht \times sex)$.

   (Δηλαδη: Για τα κορίτσια: $E(lnFEV) = \beta_0 + \beta_1 age + \beta_2 ht$, ενώ για τα αγόρια: $E(lnFEV) = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)age + (\beta_2 + \beta_5)ht$, $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$.

$$
\begin{aligned}
F &= \frac{\frac{(SSR_{full}-SSR_{red})}{3}}{MSE_{full}} \\
&= \frac{\frac{(SSR(X_1,X_2,X_3,X_4,X_5)-SSR(X_1,X_2))}{3}}{MSE_{full}} \\
&= \frac{\frac{SSR(X_3,X_4,X_5|X_1,X_2)}{3}}{MSE_{full}} \\
&= \frac{(0.1175+0.0011+0.0015)/3}{0.0204} = 1.962418
\end{aligned}
$$

$P-value = P(F_{3,583} > 1.962418) > 0.10$ Δεν απορρίπτω τη μηδενική υπόθεση καθότι $p-value > \alpha = 0.05$. Μπορώ δηλαδή να παραλείψω το φύλο.

b. (7 μονάδες) Θεωρήσετε τη μηδενική υπόθεση που λέει πως το φύλο δεν επηρεάζει ούτε την επίδραση της ηλικίας, ούτε την επίδραση του ύψους στη τιμή της σπιρομέτρησης ενός ατόμου έναντι της δίπλευρης εναλλακτικής. **Γράψτε** την τιμή της σ.σ.ε., προσέγγιση του p-value, συμπέρασμα σε απλά ελληνικά.
$H_0: \beta_4 = \beta_5 = 0.$

$$
\begin{aligned}
F &= \frac{\frac{(SSR_{full}-SSR_{red})}{2}}{MSE_{full}} \\
&= \frac{\frac{(SSR(X_1,X_2,X_3,X_4,X_5)-SSR(X_1,X_2,X_3))}{2}}{MSE_{full}} \\
&= \frac{\frac{SSR(X_4,X_5|X_1,X_2,X_3)}{2}}{MSE_{full}} \\
&= \frac{(0.0011+0.0015)/2}{0.0204} = 0.06372
\end{aligned}
$$

$P-value = P(F_{2,583} > 0.06372) > 0.10$ Δεν απορρίπτω τη μηδενική υπόθεση καθότι $p-value > \alpha = 0.05$. Μπορώ δηλαδή να παραλείψω τις αλληλεπιδράσεις με το φύλο.

c. (10 μονάδες) Βάσει του πλήρους μοντέλου **Γράψτε** το (προσεγγιστικό) 98% Δ.Ε. για την μέση διαφορά στη τιμή της σπιρομέτρησης μεταξύ 12-χρονων αγοριών με ύψος 59 ίντσες και 12-χρονων κοριτσιών με το ίδιο ύψος.
Για τα κορίτσια: $E(lnFEV) = \beta_0 + 12\beta_1 + 59\beta_2$.
Για τα αγόρια: $E(lnFEV) = (\beta_0+\beta_3)+12(\beta_1+\beta_4)+59(\beta_2+\beta_5)$.
Θέλω 98% Δ.Ε. για $\delta = \beta_3 + 12\beta_4 + 59\beta_5$.

1ος Τρόπος: Εκτίμηση $\hat{\delta} = \hat{\beta}_3 + 12\hat{\beta}_4 + 59\hat{\beta}_5 = -0.007 + 12(-0.00263) + 59(0.00101) = 0.0210$

$$\begin{aligned} Var(\hat{\delta}) &= \\ Var(\hat{\beta}_3 + 12\hat{\beta}_4 + 59\hat{\beta}_5) &= Var(\hat{\beta}_3) + 144Var(\hat{\beta}_4) + 3481Var(\hat{\beta}_5) \\ &+ 24Cov(\hat{\beta}_3, \hat{\beta}_4) + 118Cov(\hat{\beta}_3, \hat{\beta}_5) + 1416Cov(\hat{\beta}_4, \hat{\beta}_5) \\ &= \sigma^2[1.50046 + 144(0.0026919) + 3481(0.0006869) \\ &+ 24(0.0403152) + 118(-0.0311044) + 1416(-0.0010898)] \\ &= \sigma^2(0.0332813) \end{aligned}$$

Άρα $\hat{Var}(\hat{\beta}_3 + 12\hat{\beta}_4 + 59\hat{\beta}_5) = 0.0332813 \times 0.0204 = 0.00067893852$ και $s_{\hat{\delta}} = 0.026056$. Το (προσεγγιστικό) 98% Δ.Ε. για το δ είναι:

$$\begin{aligned} \hat{\delta} &\pm t_{583,0.01} s_{\hat{\delta}} \\ 0.0210 &\pm 2.33(0.026056) \\ 0.0210 &\pm 0.061 \\ (-0.04, &0.082) \end{aligned}$$

2ος Τρόπος: Παρατηρώ πως $\delta = \beta_3 + 12\beta_4 + 59\beta_5 = \boldsymbol{l}^T\boldsymbol{\beta}$ με $\boldsymbol{l}^T = (0\ 0\ 0\ 1\ 12\ 59)$. Έχω

$$\begin{aligned} \hat{\delta} &= \boldsymbol{l}^T\hat{\boldsymbol{\beta}} \\ Var(\hat{\delta}) &= Var(\boldsymbol{l}^T\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{l}^T Var(\hat{\boldsymbol{\beta}})\boldsymbol{l} \\ &= \sigma^2\boldsymbol{l}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{l} \\ &= \sigma^2(1\ 12\ 59)\boldsymbol{W}(1\ 12\ 59)^T \end{aligned}$$

όπου $\boldsymbol{W}$ είναι ο 3×3 κάτω υποπίνακας του $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Κάνω τις πράξεις και καταλήγω στο ίδιο συμπέρασμα.

d. (*****) Ελέγξτε τη μηδενική υπόθεση που λέει πως δεν χρειάζεται να ληφθεί το φύλο υπόψη, όταν το πλήρες μοντέλο είναι $E(lnFEV) = \beta_0 + \beta_1 age + \beta_2 ht + \beta_3 sex$

$H_0 : \beta_3 = 0$. (Δηλαδή το περιορισμένο μοντέλο είνα $E(lnFEV) = \beta_0 + \beta_1 age + \beta_2 ht$.).

$$
\begin{aligned}
F &= \frac{\frac{(SSR_{full}-SSR_{red})}{1}}{MSE_{full}} \\
&= \frac{\frac{(SSR(X_1,X_2,X_3)-SSR(X_1,X_2))}{1}}{MSE_{full}} \\
&= \frac{\frac{SSR(X_3|X_1,X_2)}{1}}{MSE_{full}} \\
&= \frac{(0.1175)/1}{(11.8793+0.0015+0.0011)/585} = 5.788
\end{aligned}
$$

$P-value = P(F_{1,585} > 5.788)$, κι έχω $0.01 < P-value < 0.025$ και απορρίπτω τη μηδενική υπόθεση (πρέπει να προσθέσω το φύλο ως κύριο παράγοντα).

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**.

Output για 3η 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 4η άσκηση

NON-SMOKERS

Regression Analysis: ln(FEV) versus age; ht

Regression Equation

ln(FEV) = -1,9333 + 0,02469 age + 0,04267 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 5η 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