

Άσκηση 2

Ο επόμενος πίνακας δίνει τους βαθμούς 20 φοιτητών (X_i) στις εισαγωγικές εξετάσεις ενός колле-γίου και τους αντίστοιχους βαθμούς τους (Y_i) στο τέλος της πρώτης χρονιάς φοίτησης στο συγκεκριμένο колέγιο.

i	β. 1 ^{ου} έτους (Y_i)	β. εισαγωγής (X_i)
1	3.1	5.5
2	2.3	4.8
3	3.0	4.7
4	1.9	3.9
5	2.5	4.5
6	3.7	6.2
7	3.4	6.0
8	2.6	5.2
9	2.8	4.7
10	1.6	4.3
11	2.0	4.9
12	2.9	5.4
13	2.3	5.0
14	3.2	6.3
15	1.8	4.6
16	1.4	4.3
17	2.0	5.0
18	3.8	5.9
19	2.2	4.1
20	1.5	4.7

Να γίνει το διάγραμμα διασποράς (scatterplot) μεταξύ των X , Y .

Να κάνετε μελέτη του μοντέλου $Y = \beta_0 + \beta_1 X + \varepsilon$. Συγκεκριμένα:

α. Δώστε την φυσική ερμηνεία των β_0 , β_1 , R^2 .

β. Κάνετε τους ελέγχους σε ε.σ. $\alpha=5\%$ και $\alpha=1\%$:

$H_0: \beta_1=0$, $H_1: \beta_1 \neq 0$, και $H_0: \beta_1=0$, $H_1: \beta_1 > 0$

$H_0: \beta_0=0$, $H_1: \beta_0 \neq 0$, και $H_0: \beta_0=0$, $H_1: \beta_0 > 0$

$H_0: \beta_1=0.8$, $H_1: \beta_1 \neq 0.8$, $H_0: \beta_0 \leq 0.8$, $H_1: \beta_1 > 0.8$, $H_0: \beta_1 \geq 0.8$, $H_1: \beta_1 < 0.8$

$H_0: \beta_0=-1$, $H_1: \beta_1 \neq -1$,

και σχολιάστε τα αποτελέσματα.

γ. Δώστε δ.ε. με σ.ε. 95% και 99% για τα β_0 , β_1 και σχολιάστε τα αποτελέσματα.

δ. Κάνετε το διάγραμμα διασποράς των δεδομένων (X, Y) μαζί με την προσαρμοσμένη ευθεία και τις ζώνες εμπιστοσύνης (για ατομική και μέση πρόβλεψη) με σ.ε. 95% και 99%.

ε. Ποιο βαθμό αναμένεται να λάβει στο 1^ο έτος σπουδών του κάποιος που έχει βαθμό εισαγωγής 5; (να γίνει σημειακή εκτίμηση και να δοθεί το δ.ε. με σ.ε. 95% και 99%). Να δοθεί επίσης το δ.ε. συντελεστού 95% για την ατομική πρόβλεψη.

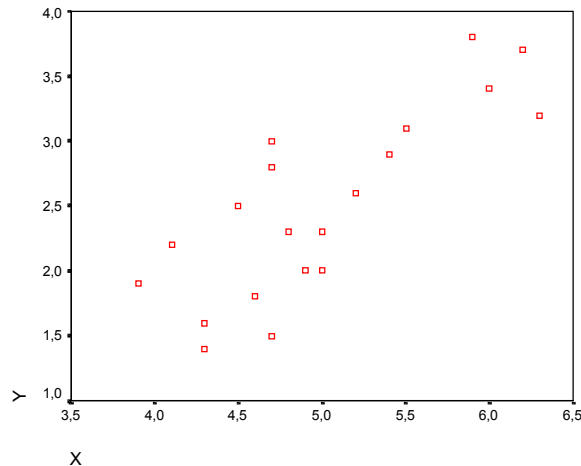
Ποιος είναι περίπου ο βαθμός εισαγωγής ενός φοιτητή που είχε βαθμό 1^{ου} έτους 3.5;

στ. Σε ποίο σημείο X βρίσκουμε το μικρότερο δ.ε. για την μέση πρόβλεψη και γιατί; ποιο είναι εδώ αυτό το δ.ε.;

ζ. Να κατασκευάσετε τον πίνακα ανάλυσης διασποράς (ANOVA) και να κάνετε τον έλεγχο του μοντέλου μέσω του F-τεστ.

Απαντήσεις.

Να γίνει το διάγραμμα διασποράς (scatterplot) μεταξύ των X , Y .
(/Graphs/Scatter ή /Graphs/Interactive/Scatterplot)



Να κάνετε μελέτη του μοντέλου $Y = \beta_0 + \beta_1 X + \varepsilon$.
(/Analyze/Regression/Linear, +/statistics/conf. intervals)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,809	,654	,635	,4350

a Predictors: (Constant), X

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6,434	1	6,434	33,998	,000
	Residual	3,406	18	,189		
	Total	9,840	19			

a Predictors: (Constant), X

b Dependent Variable: Y

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
Model		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1,700	,727		-2,338	,031	-3,226	-,173
	X	,840	,144	,809	5,831	,000	,537	1,143

a Dependent Variable: Y

α. Δώστε την φυσική ερμηνεία των β_0 , β_1 , R^2 .

Το μοντέλο που χρησιμοποιούμε είναι: $Y = \beta_0 + \beta_1 X + \varepsilon$ και άρα $E(Y) = \beta_0 + \beta_1 X$. Συνεπώς,

$$\text{Μέσος Βαθμός 1}^{\text{ου}} \text{ έτους} = \beta_0 + \beta_1 \times \text{Βαθμός εισαγωγής}$$

Επομένως, το β_1 μπορεί να ερμηνευθεί ως η (μέση) αύξηση του Βαθμού 1^{ου} έτους όταν αυξηθεί ο Βαθμός εισαγωγής κατά μία μονάδα. Αντίστοιχα, το β_0 μπορεί να θεωρηθεί ως ο (μέσος) Βαθμός 1^{ου} έτους για κάποιον που (θεωρητικά) είχε μηδενικό βαθμό εισαγωγής. Τέλος, το R^2 μπορεί να θεωρηθεί ως το ποσοστό της μεταβλητότητας των Βαθμών του 1^{ου} έτους που ερμηνεύεται από το μοντέλο $Y = \beta_0 + \beta_1 X + \varepsilon$.

β. Κάνετε τους ελέγχους σε ε.σ. $\alpha=5\%$ και $\alpha=1\%$:

$H_0: \beta_1=0$, $H_1: \beta_1 \neq 0$,

Για τον συγκεκριμένο έλεγχο λαμβάνουμε από τον πίνακα των Coefficients ότι το αντίστοιχο p-value είναι (περιοχή απόρριψης: $|T| > |T^*|$)

$$p_0 = P(|T| > |T^*|) = P(|T| > 5.831) \approx 0.000$$

οπότε απορρίπτουμε την υπόθεση H_0 σε ε.σ. $\alpha=5\%$ και $\alpha=1\%$ ($0.000 < \alpha = 0.05, 0.01$).

$H_0: \beta_1=0, H_1: \beta_1>0$

Το αντίστοιχο p-value είναι (περιοχή απόρριψης: $T > T^*$)

$$P(T > T^*) = P(T > 5.831) = \frac{1}{2} P(|T| > 5.831) = \frac{1}{2} p_0 \approx 0.000$$

οπότε απορρίπτουμε την υπόθεση H_0 σε ε.σ. $\alpha = 5\%$ και $\alpha = 1\%$ ($0.000 < \alpha = 0.05, 0.01$).

Εναλλακτικά, αν $\alpha = 5\%$ απορρίπτουμε όταν

$$T^* > t_{n-2}(\alpha) = t_{18}(0.05) = 1.73$$

και αν $\alpha = 1\%$ όταν

$$T^* > t_{n-2}(\alpha) = t_{18}(0.01) = 2.55$$

Επειδή $T^* = 5,831$ απορρίπτουμε και στις δύο περιπτώσεις.

$H_0: \beta_0=0, H_1: \beta_0 \neq 0$,

Για τον συγκεκριμένο έλεγχο λαμβάνουμε από τον πίνακα των Coefficients ότι το αντίστοιχο p-value είναι (περιοχή απόρριψης: $|T| > |T^*|$)

$$p_0 = P(|T| > |T^*|) = P(|T| > 2.338) \approx 0.031$$

οπότε απορρίπτουμε την υπόθεση H_0 : σε ε.σ. $\alpha = 5\%$ ($0.031 < \alpha = 0.05$) ενώ δεν μπορούμε να απορρίψουμε την H_0 όταν το ε.σ. είναι $\alpha = 1\%$ ($0.031 > \alpha = 0.01$).

$H_0: \beta_0=0, H_1: \beta_0>0$

Το αντίστοιχο p-value είναι (περιοχή απόρριψης: $T > T^*$)

$$P(T > T^*) = P(T > -2.338) = 1 - \frac{1}{2} P(|T| > 2,338) = 1 - \frac{1}{2} p_0 \approx 1 - \frac{1}{2} 0.031 = 0.9845$$

οπότε δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. $\alpha = 5\%$ και $\alpha = 1\%$.

Εναλλακτικά, αν $\alpha = 5\%$ απορρίπτουμε όταν

$$T^* > t_{n-2}(\alpha) = t_{18}(0.05) = 1.73$$

και αν $\alpha = 1\%$ όταν

$$T^* > t_{n-2}(\alpha) = t_{18}(0.01) = 2.55$$

Επειδή $T^* = -2.338$ δεν μπορούμε να απορρίψουμε και στις δύο περιπτώσεις.

$H_0: \beta_1=0.8, H_1: \beta_1 \neq 0.8$,

Ο έλεγχος αυτός δεν γίνεται αυτόματα από το πακέτο. Για το λόγο αυτό θα πρέπει να μετασχηματίσουμε τα δεδομένα ώστε να καταλήξουμε σε έλεγχο της μορφής $\beta'_1 = 0$. Συγκεκριμένα λαμβάνοντας $\beta'_1 = \beta_1 - 0.8$ προκύπτει το μοντέλο

$$Y = \beta_0 + \beta_1 X + \varepsilon \Leftrightarrow Y = \beta_0 + (\beta'_1 + 0.8)X + \varepsilon \Leftrightarrow Y - 0.8X = \beta_0 + \beta'_1 X + \varepsilon$$

Θέτοντας λοιπόν $Y_1 = Y - 0.8X$ (/Transform/Compute) και εφαρμόζοντας το μοντέλο $Y_1 = \beta_0 + \beta'_1 X + \varepsilon$ (/Analyze/Regression/ Linear) λαμβάνουμε τον πίνακα

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	-1,700	,727		-2,338	,031
	X	3,991E-02	,144	,065	,277	,785

a Dependent Variable: Y1

Παρατηρούμε ότι το αντίστοιχο p-value είναι

$$p_0 = P(|T| > |T^*|) = P(|T| > 0,277) \approx 0.785$$

οπότε δεν μπορούμε να απορρίψουμε την $H_0 : \beta'_1 = 0 \Leftrightarrow \beta_1 = 0.8$ σε ε.σ. $\alpha = 5\%$ και $\alpha = 1\%$.

$H_0: \beta_1 \leq 0.8, H_1: \beta_1 > 0.8,$

Το αντίστοιχο p-value είναι (περιοχή απόρριψης: $T > T^*$)

$$P(T > T^*) = P(T > 0.277) = \frac{1}{2} P(|T| > 0.277) \approx \frac{1}{2} 0.785 = 0.3925$$

οπότε δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. $\alpha = 5\%$ και $\alpha = 1\%$.

Εναλλακτικά, αν $\alpha = 5\%$ απορρίπτουμε όταν $T^* > t_{n-2}(\alpha) = t_{18}(0.05) = 1.73$ και αν $\alpha = 1\%$ όταν $T^* > t_{n-2}(\alpha) = t_{18}(0.01) = 2.55$. Επειδή $T^* = 0.277$ δεν μπορούμε να απορρίψουμε και στις δύο περιπτώσεις.

$H_0: \beta_1 \geq 0.8, H_1: \beta_1 < 0.8$

Το αντίστοιχο p-value είναι (περιοχή απόρριψης: $T < T^*$)

$$P(T < T^*) = P(T < 0.277) = 1 - \frac{1}{2} P(|T| > 0.277) = 1 - \frac{1}{2} p_0 \approx 1 - \frac{1}{2} 0.785 = 0.6075$$

οπότε δεν μπορούμε να απορρίψουμε την H_0 σε ε.σ. $\alpha = 5\%$ και $\alpha = 1\%$.

Εναλλακτικά, αν $\alpha = 5\%$ απορρίπτουμε όταν

$$T^* < -t_{n-2}(\alpha) = -t_{18}(0.05) = -1.73$$

και αν $\alpha = 1\%$ όταν

$$T^* < -t_{n-2}(\alpha) = -t_{18}(0.01) = -2.55$$

Επειδή $T^* = 0.277$ δεν μπορούμε να απορρίψουμε και στις δύο περιπτώσεις.

$H_0: \beta_0 = -1, H_1: \beta_0 \neq -1,$

Θέτουμε $\beta'_0 = \beta_0 + 1$ όποτε προκύπτει το μοντέλο

$$Y = \beta_0 + \beta_1 X + \varepsilon \Leftrightarrow Y = \beta'_0 - 1 + \beta_1 X + \varepsilon \Leftrightarrow Y + 1 = \beta'_0 + \beta_1 X + \varepsilon$$

Θέτοντας λοιπόν $Y_1 = Y + 1$ (/Transform/Compute) και εφαρμόζοντας το μοντέλο $Y_1 = \beta'_0 + \beta_1 X + \varepsilon$

(/Analyze/Regression/ Linear) λαμβάνουμε τον πίνακα

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	-,700	,727		-,963	,349
	X	,840	,144	,809	5,831	,000

a Dependent Variable: Y1

Παρατηρούμε ότι το αντίστοιχο p-value είναι 0.349 οπότε δεν μπορούμε να απορρίψουμε την $H_0 : \beta'_0 = 0 \Leftrightarrow \beta_0 = -1$ σε ε.σ. $\alpha = 5\%$ και $\alpha = 1\%$.

γ. Δώστε δ.ε. με σ.ε. 95% και 99% για τα β_0, β_1 .

Τα δ.ε. 95% δίνονται αυτόματα από το πακέτο (/Analyze/Regression/Linear,+/statistics/conf. intervals√) και περιέχονται στον πίνακα των coefficients:

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
Model		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1,700	,727		-2,338	,031	-3,226	-,173
	X	,840	,144	,809	5,831	,000	,537	1,143

a Dependent Variable: Y

Άρα ένα δ.ε. 95% για το β_0 είναι το $(-3.226, -0.173)$ και ένα δ.ε. 95% για το β_1 είναι το $(0.537, 1.143)$.

Παρατηρούμε ότι τα δ.ε. δεν περιέχουν το 0 και συνεπώς μπορούμε να απορρίψουμε τις υποθέσεις

$H_0:\beta_0 = 0$ και $H_0:\beta_1 = 0$ έναντι των αμφίπλευρων $H_1:\beta_0 \neq 0$ και $H_1:\beta_1 \neq 0$ αντίστοιχα σε ε.σ. 5% (όπως έχουμε ήδη κάνει στο προηγούμενο ερώτημα).

Τα αντίστοιχα δ.ε. 1% δεν δίνονται αυτόματα από το πακέτο και άρα θα τα υπολογίσουμε μέσω των τύπων

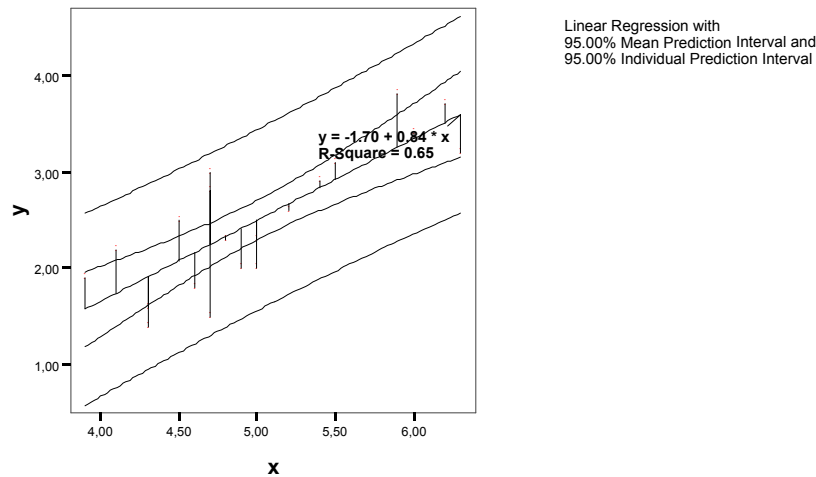
$$\hat{\beta}_0 \pm s(\hat{\beta}_0) t_{n-2}(a/2) = -1.7 \pm 0.727 \cdot t_{18}(0.005) = -1.7 \pm 0.727 \cdot 2.878 = (-3.792, 0.392)$$

$$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_{n-2}(a/2) = 0.840 \pm 0.144 \cdot t_{18}(0.005) = 0.840 \pm 0.144 \cdot 2.878 = (0.425, 1.254)$$

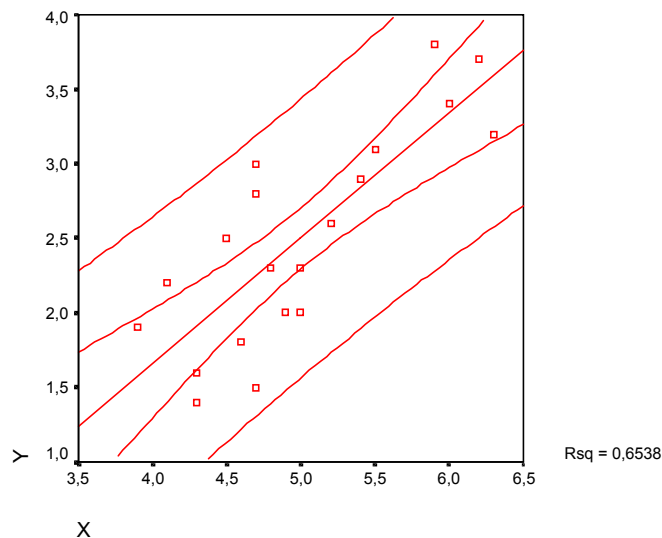
(π.χ. χρησιμοποιούμε την function $IDF.T(0.995, 18) = t_{18}(0.005)$ στο Transform/compute). Παρατηρούμε ότι το δ.ε. για το β_1 δεν περιέχει το 0 και συνεπώς μπορούμε να απορρίψουμε την υπόθεση $H_0:\beta_1 = 0$ έναντι της $H_1:\beta_1 \neq 0$ σε ε.σ. 1%. Αντίθετα, το δ.ε. για το β_0 περιέχει το 0 και συνεπώς δεν μπορούμε να απορρίψουμε την υπόθεση $H_0:\beta_0 = 0$ έναντι της $H_1:\beta_0 \neq 0$ σε ε.σ. 1%.

δ. Κάνετε το διάγραμμα διασποράς των δεδομένων (X,Y) μαζί με την προσαρμοσμένη ευθεία και τις ζώνες εμπιστοσύνης (για ατομική και μέση πρόβλεψη) με σ.ε. 95% και 99%.

α-τρόπος: (Graphs/interactive/scatterplot , Fit: regression, conf. Interval 95%)



β-τρόπος: Graphs/Scatter/simple, και κάνοντας διπλό κλικ επάνω στο σχήμα που θα δημιουργηθεί, ανοίγει ο SPSS Chart editor του σχήματος και επιλέγουμε Chart/options, Fit line:total, Fit options:Linear regression, Regression prediction lines: Mean, Individual, 95%)



Όμοια εργαζόμαστε και για την περίπτωση που έχουμε σ.ε. 99%.

ε. Ποιο βαθμό αναμένεται να λάβει στο 1^ο έτος σπουδών του κάποιος που έχει βαθμό εισαγωγής 5; (να γίνει σημειακή εκτίμηση και να δοθεί το δ.ε. με σ.ε. 95% και 99%). Να δοθεί επίσης το δ.ε. συντελεστού 95% για την ατομική πρόβλεψη. Ποιος είναι περίπου ο βαθμός εισαγωγής ενός φοιτητή που είχε βαθμό 1^{ου} έτους 3.5;

Εδώ ζητείται εκτίμηση του $E(Y) = \beta_0 + \beta_1 X$ όταν $X = 5$, δηλαδή του $\beta_0 + 5\beta_1$. Η σημειακή εκτίμηση του θα είναι

$$\hat{\beta}_0 + 5\hat{\beta}_1 = -1.7 + 5 \cdot 0.840 = 2.5.$$

Για να πάρουμε δ.ε. 95% για τη μέση πρόβλεψη εισάγουμε το 5 στην 21^η γραμμή της στήλης του X στον SPSS data editor (αφήνοντας την 21^η γραμμή στη στήλη του Y κενή). Στη συνέχεια εκτελούμε τη διαδικασία της παλινδρόμησης /Analyze/Regression/Linear επιλέγοντας στο save τα Unstandardized predicted values ($\hat{Y}_o = \hat{\beta}_0 + \hat{\beta}_1 X_o$), Prediction Intervals (Mean & Individual, 95%). Στην 21^η γραμμή λαμβάνονται τα αποτελέσματα:

σημειακή εκτίμηση: 2,50000 (όπως έχουμε ήδη βρεί και παραπάνω)

δ.ε. 95% για την μέση πρόβλεψη: (2.29564, 2.70436)

δ.ε. 95% για την ατομική πρόβλεψη: (1.56350, 3.43650)

Για να πάρουμε δ.ε. 99% για τη μέση πρόβλεψη εργαζόμαστε ανάλογα. Στην 21^η γραμμή λαμβάνεται το δ.ε. (2.22001, 2.77999).

Αν ο φοιτητής είχε βαθμό 1^{ου} έτους 3.5, ο βαθμός εισαγωγής εκτιμάται από το

$$X = \frac{Y - \hat{\beta}_0}{\hat{\beta}_1} = \frac{3.5 + 1.7}{0.840} \approx 6.19.$$

στ. Σε ποίο σημείο X βρίσκουμε το μικρότερο δ.ε. για την μέση πρόβλεψη και γιατί; ποιο είναι εδώ αυτό το δ.ε.;

Το δ.ε. για την μέση πρόβλεψη έχει τη μορφή

$$\hat{Y}_o \pm s(\hat{Y}_o) \cdot t_{n-2}(a/2) = \hat{\beta}_0 + \hat{\beta}_1 X_o \pm s \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{XX}}} \cdot t_{n-2}(a/2),$$

με εύρος

$$2s \sqrt{\frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{XX}}} \cdot t_{n-2}(a/2)$$

το οποίο προφανώς θα είναι ελάχιστο όταν $X_o = \bar{X}$. Από την διαδικασία Descriptive Statistics βρίσκουμε ότι $\bar{X} = 5$. Το αντίστοιχο δ.ε. έχει υπολογιστεί στο προηγούμενο ερώτημα.

ζ. Να κατασκευάσετε τον πίνακα ανάλυσης διασποράς (ANOVA) και να κάνετε τον έλεγχο του μοντέλου μέσω του F-τεστ.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6,434	1	6,434	33,998	,000
	Residual	3,406	18	,189		
	Total	9,840	19			

a Predictors: (Constant), X

b Dependent Variable: Y

Ο έλεγχος του μοντέλου μέσω του F -test μας οδηγεί στην απόρριψη της $H_0: \beta_1=0$ διότι το αντίστοιχο p -value

$$p\text{-value} = P(F > F^* \mid F \sim F_{1,n-2}) = P(F > 33,998 \mid F \sim F_{1,18})$$

είναι ίσο με 0.000.