ΧΑΤΖΟΠΟΥΛΟΣ ΓΕΡΑΣΙΜΟΣ

Στον πίνακα που ακολουθεί δίνονται οι μετρήσεις τεσσάρων δεικτών από αιματολογική εξέταση ενός δείγματος 20 ασθενών ενός ενδοκρινολογικού τμήματος. Οι μετρήσεις αυτές αφορούν τους κάτωθι δείκτες:

HB: αιμοσφαιρίνη (g/100ml)
HT: Αιματοκρίτης (%)
MCV: Μέσος όγκος ερυθρών (κυβ. μικρά)
MCH: Μέση περιεκτικότητα ερυθρών (μμg)

HB	HT	MCV	MCH
13,1	36	77	20
12,6	32	67	19
15,8	38	98	28
16,0	39	66	28
17,3	38	73	29
11,8	34	65	25
14,7	41	59	25
18,2	48	90	27
14,5	33	56	23
15,5	36	68	25
16,2	38	77	28
17,8	40	79	29
21,0	56	110	25
19,4	46	102	30
16,4	40	88	28
14,2	42	76	29
12,9	33	65	24
15,6	37	87	28
11,9	32	64	23
10,2	37	60	22

α) Σκοπός μας είναι να κάνουμε ένα σύντομο report και να βρούμε το καλύτερο μοντέλο για την HB μεταβλητή μας και ποσό επηρεάζεται από τις άλλες τρεις

1)Ας ξεκινήσω μεταφέροντας τις μεταβλητές μου στην R και μετα βρίσκοντας τα correlation μεταξύ όλων των μεταβλητών

```
datatest <- read.csv("C:/Users/gerry/R/data.txt", sep="")
```

```
HB <- data[,1]
HT <- data[,2]
MCH <- data[,3]
MCV <- data[,4]
```

```
data<-data.frame(HB,HT,MCH,MCV)

cor(data)
          HB        HT       MCH       MCV
HB  1.0000000 0.7964686 0.6387237 0.7627754
HT  0.7964686 1.0000000 0.4360949 0.7478197
MCH 0.6387237 0.4360949 1.0000000 0.4752019
MCV 0.7627754 0.7478197 0.4752019 1.0000000
```

Ας ξεκινήσουμε δημιουργώντας το γραμμικό μοντέλο της HB με την HT

```
y2<-lm(HB~HT)
 summary(y2)

Call:
lm(formula = HB ~ HT)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3916 -0.9173  0.2613  1.2491  2.3399

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95473    2.58674   0.369    0.716
HT           0.36856    0.06595   5.588 2.65e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.691 on 18 degrees of freedom
Multiple R-squared:  0.6344,    Adjusted R-squared:  0.614
F-statistic: 31.23 on 1 and 18 DF,  p-value: 2.647e-05
```

(β) Να <u>εξηγηθούν</u> οι εκτιμήσεις των παραμέτρων του μοντέλου.

Από τον πίνακα Coefficients, βλέπουμε ότι το μοντέλο μας παίρνει την εξής μορφή:

$\hat{Y} = 0.955 + 0.369 \cdot X1$

Για το $\beta 0$: Βάσει της εκτίμησης του σταθερού όρου του μοντέλου, η μέση τιμή της αιμοσφαιρίνης (HB) είναι 0.955, όταν το ποσοστό του αιματοκρίτη (HT) είναι 0 στο αίμα του ασθενή.

Για το $\beta 1$: Όταν αυξάνεται κατά μία μονάδα η τιμή του αιματοκρίτη (HT), τότε αναμένουμε αύξηση της τιμής της αιμοσφαιρίνης (HB) κατά 0.369 μονάδες

(γ) Να ελεγχθεί η καταλληλότητα του μοντέλου.

```
anova(y2)

Analysis of Variance Table

Response: HB
           Df  Sum Sq Mean Sq F value    Pr(>F)
HT          1  89.273  89.273  31.229 2.647e-05 ***
Residuals 18  51.456   2.859
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Μιας και έχουμε το απλό γραμμικό μοντέλο, τα t-test και F-test είναι ισάξια, μιας και ελέγχουν μόνο μία μεταβλητή. Θα κάνω λοιπόν το F-test, μιας και έχω ήδη υπολογίσει την τιμή της F από τον πίνακα ANOVA.

Για να δω αν το μοντέλο μου είναι κατάλληλο, αυτό που θα κάνω είναι να ελέγξω τις εξής υποθέσεις:
$H0$: $\beta 1 = 0$
$H1$: $\beta 1 \neq 0$

Σε επίπεδο στατιστικής σημαντικότητας, έστω $\alpha = 0.05$.
Η μηδενική υπόθεση δηλώνει ότι δεν υπάρχει γραμμική σχέση μεταξύ των Χ και Υ, ενώ η εναλλακτική δηλώνει το ακριβώς αντίθετο.

Έχουμε ήδη υπολογίσει την F-statistic από το Α) F-statistic=31.22

το μόνο που μας μένει είναι να συγκρίνουμε αυτήν την τιμή με:

$= F(1,19,0.975) = 5.922$
$(k,n-k-1,1-2)$
Παρατηρούμε λοιπόν ότι:
$F(1,19) > F(1,19,0.975) \Longrightarrow 31.225 > 5.922$

Επομένως σε επίπεδο στατιστικής σημαντικότητας 5%,
απορρίπτουμε την μηδενική υπόθεση ($H0$), δηλώνοντας έτσι ότι
τελικά υπάρχουν ισχυρές ενδείξεις για γραμμική σχέση μεταξύ των ΗΒ και ΗΤ.

(δ) Να υπολογιστούν τα διαστήματα εμπιστοσύνης βαθμού 95%

για τις παραμέτρους του μοντέλου και να εξηγηθούν.

```
confint(y2)
```

	2.5 %	97.5 %
(Intercept)	-4.479817	6.3892816
HT	0.230002	0.5071252

Εξήγηση διαστήματος εμπιστοσύνης B_0: Με πιθανότητα 95%, ο σταθερός όρος του μοντέλου παλινδρόμησης στον πληθυσμό ($\beta 0$), από τον οποίο προήλθαν οι παρατηρήσεις μας, θα είναι κάπου ανάμεσα στο -4.480287 και το 6.390287. Επειδή μέσα στο διάστημα αυτό περιλαμβάνεται η τιμή 0, μπορούμε να ισχυριστούμε ότι ο σταθερός όρος δεν είναι στατιστικά σημαντικός.

Εξήγηση διαστήματος εμπιστοσύνης B_1: Στο 95% των ασθενών ολόκληρου του πληθυσμού, από μία αύξηση του αιματοκρίτη (HT) κατά μία μονάδα, αναμένεται αύξηση της αιμοσφαιρίνης (HB) κάπου ανάμεσα σε 0.230334 και 0.506666 μονάδες. Εδώ μπορούμε να πούμε ότι ο όρος $\beta 1$ είναι στατιστικά σημαντικός, μιας και το διάστημα εμπιστοσύνης του δεν περιλαμβάνει την τιμή 0.

(ε) Να υπολογιστεί και να εξηγηθεί διαστημα εμπιστοσύνης βαθμου 95% για την μέση τιμή της HB όταν η HT ισούται με 30.

```
predict.lm(y2,newdata=data.frame(HT=30),interval="confidence")
```

	fit	lwr	upr
1	12.01164	10.55641	13.46687

Αυτό, θα περιέχει την αναμενόμενη τιμή της αιμοσφαιρίνης για δεδομένη τιμή αιματοκρίτη με βαθμό εμπιστοσύνης 0.95.Δηλαδή, είμαστε 95% σίγουροι ότι η πραγματική τιμή της μέσης (ή αναμενόμενης) τιμής της Y (αιμοσφαιρίνης) όταν η X (ο αιματοκρίτης) είναι 30 θα βρίσκεται στο διάστημα [10.421, 13.613]. Με άλλα λόγια εάν κατασκευάσουμε πολλές φορές διάστημα εμπιστοσύνης για την αναμενόμενη τιμή της αιμοσφαιρίνης όταν ο αιματοκρίτης είναι 30, τότε το ποσοστό αυτών που θα περιέχουν αυτήν την τιμή θα προσεγγίζει το 95% (όσο αυξάνεται ο αριθμός των φορών που κατασκευάζουμε διαστήματα εμπιστοσύνης) και θα βρίσκεται μεταξύ του διαστήματος [10.421, 13.613]

(στ) Να υπολογιστεί και να εξηγηθεί το 95% διάστημα πρόβλεψης για την τιμή της HB όταν η HT ισούται με 30.

```
predict.lm(y2,newdata=data.frame(HT=30),interval="predict")
```

```
       fit      lwr      upr
1 12.01164 8.172958 15.85032
```

Επομένως, είμαστε κατά 95% σίγουροι οτι με επίπεδα αιματοκρίτη (HT) ίσα με 30 μονάδες, εκτιμάται ότι το επίπεδο αιμοσφαιρίνης (HB) θα βρίσκεται κάπου ανάμεσα στις 8.2 και 15.9 μονάδες.

ζ)Ας περασουμε τωρα στην κατασκευη του μοντελου παλινδρομισης με εξαρτημενη παλι αυτην που μας ενδιαφερει δηλαδη την HB αλλα αυτην την φορα με ανεξαρτητες τις HT και την MCH

```
y3<-lm(HB~HT+MCH)
 summary(y3)
```

```
Call:
lm(formula = HB ~ HT + MCH)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3603 -0.2464  0.1366  0.9610  1.8137

Coefficients:
             Estimate     Std. Error     t value     Pr(>|t|)
(Intercept) -4.20716     2.99197      -1.406     0.177695
HT           0.29595     0.06369       4.647     0.000231 ***
MCH          0.30987     0.11853       2.614     0.018135 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.469 on 17 degrees of freedom
Multiple R-squared:  0.7392,       Adjusted R-squared:  0.7085
F-statistic: 24.09 on 2 and 17 DF,  p-value: 1.093e-05
```

Από τον πίνακα Coefficients, βλέπουμε ότι το μοντέλο μας παίρνει την εξής μορφή:

$\hat{Y} = -4.207 + 0.296 \cdot X1 + 0.310 \cdot X2$

Για το $\beta 0$: Βάσει της εκτίμησης του σταθερού όρου του μοντέλου, η μέση τιμή της αιμοσφαιρίνης (HB) είναι −4.207, όταν το ποσοστό του αιματοκρίτη (HT), καθώς και η μέση περιεκτικότητα ερυθρού (MCH) είναι 0 στο αίμα του ασθενή.

Για το $\beta 1$: Όταν αυξάνεται κατά μία μονάδα η τιμή του αιματοκρίτη (HT), ενώ η μέση περιεκτικότητα ερυθρών (MCH) του ασθενή παραμένει σταθερή, τότε αναμένουμε αύξηση της τιμής της αιμοσφαιρίνης (HB) κατά 0.296 μονάδες.

Για το $\beta 2$: Όταν αυξάνεται κατά μία μονάδα η μέση περιεκτικότητα ερυθρών (MCH), ενώ η τιμή του αιματοκρίτη (HT) του ασθενή παραμένει σταθερή, τότε αναμένουμε αύξηση της τιμής της αιμοσφαιρίνης (HB) κατά 0.310 μονάδες.

Όσον αφορά τον adjusted συντελεστή προσδιορισμού του μοντέλου

Το s αποτελεί ένα κριτήριο καλής προσαρμογής (όσο μικρότερη είναι η τιμή του, τόσο καλύτερη η προσαρμογή του μοντέλου στα δεδομένα μας). Εν τούτοις έχει το βασικό μειονέκτημα ότι μετράται στις μονάδες μέτρησης της εξαρτημένης μεταβλητής μας. Έτσι λοιπόν είναι ένα μέτρο ακατάλληλο για συγκρίσεις μεταξύ διαφορετικών εφαρμογών.

Ένα άλλο μέτρο που μπορεί κανείς να σκεφτεί είναι ο συντελεστής προσδιορισμού $R2$. Όμως, ο συντελεστής γίνεται ακατάλληλος για συγκρίσεις μεταξύ μοντέλων με διαφορετικό αριθμό ανεξαρτήτων μεταβλητών, αφού κάθε φορά που θα προστίθεται μια ακόμα ανεξάρτητη μεταβλητή στο μοντέλο, η τιμή του θα αυξάνεται ανεξάρτητα από το αν η συμβολή της επιπλέον μεταβλητής είναι σημαντική ή όχι.

Επομένως, το μόνο μέτρο που θα χρησιμοποιήσουμε για να συγκρίνουμε τα δύο αυτά μοντέλα, είναι ο προσαρμοσμένος συντελεστής προσδιορισμού $\bar{R}2$ (Adjusted R- Square). Παρατηρούμε ότι η τιμή του αυξήθηκε από το 0.614(HB~HT) στο 0.709(HB~HT+MCH), πράγμα που σημαίνει ότι το δεύτερο μοντέλο είναι καλύτερο από το απλό γραμμικό, διότι ερμηνεύει το 70.9% της μεταβλητότητας του Y.

Ενας άλλος και πιο γρήγορος τρόπος για την σύγκριση του απλού γραμμικού μοντέλου με το πολλαπλό γραμμικό μοντέλο και γενικώς για την εύρεση του βέλτιστου μοντέλου χωρίς μεγάλο κόπο είναι μέσο του κριτηρίου AIC και BIC αλλά θα το εξηγήσω λίγο πιο κάτω όταν θαχω φτιάξει μοντέλο και με την 3η ανεξάρτητη μεταβλητή του

Στη συνέχεια ας φτιαξω και το μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή την HB και ανεξάρτητες και τις τρεις άλλες μεταβλητές (HT, MCV και MCH),

```
y4<-lm(HB~HT+MCH+MCV)
summary(y4)
```

```
Call:
lm(formula = HB ~ HT + MCH + MCV)

Residuals:
     Min       1Q   Median       3Q      Max
-2.87047 -0.44393  0.01625  0.75119  2.20478

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  -3.55697    2.92726    -1.215    0.2419
HT            0.21111     0.08431   2.504    0.0235 *
MCH           0.26633    0.11839    2.250    0.0389 *
MCV           0.04928    0.03344    1.474    0.1599
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.421 on 16 degrees of freedom
Multiple R-squared:  0.7704,      Adjusted R-squared:  0.7273
F-statistic: 17.89 on 3 and 16 DF,  p-value: 2.301e-05
```

(ι) Είναι αυτό το μοντέλο καταλληλότερο του προηγουμένου

Για να συγκρίνουμε τα δύο αυτά μοντέλα, θα κοιτάξουμε πάλι τον προσαρμοσμένο συντελεστή προσδιορισμού $\bar{R}2$ (Adjusted R-Square). Παρατηρούμε ότι αυξήθηκε από 0.709 στο 0.727. Πράγμα που σημαίνει ότι τώρα το νέο αυτό μοντέλο είναι καταλληλότερο του προηγούμενου, μιάς και ερμηνεύει το 72.7% της μεταβλητότητας του Y.
Οπότε με βάση διορθωμένο τον συντελεστή προσδιορισμού καταλήγουμε ότι το μοντέλο που εξηγεί βέλτιστα την μεταβλητή που μας ενδιαφέρει (HB) δηλαδή για να πάρουμε την πιο σωστή και ακριβή εξήγηση θα πραγματοποιηθεί αν βάλουμε και τις 3 μεταβλητές μας (HT,MCH,MCV) ως επεξηγηματικές στο μοντέλο.

AIC-BIC

Ας έρθουμε στο θέμα όμως που ανάφερα λίγο πιο πάνω δηλαδή ότι θα μπορούσαμε να το βρούμε το βέλτιστο μοντέλο μέσο του AIC,BIC.
Όπου μπορούμε να το πραγματοποιήσουμε και για επιβεβαίωση
Και για επαλήθευση του κριτηρίου με τον συντελεστή προσδιορισμού

Backward elimination for AIC

Όπου ξεκινά από το **γεμάτο** μοντέλο μας και αφαιρει κάθε φορά μια μεταβλητή και το ξανασυμβαίνει μέχρι να φτάσει στο μοντέλο y~1 δηλαδή να μην υπάρχει καμία επεξηγηματική μεταβλητή.
Όποιος συνδυασμός έχει το μικρότερο AIC σημαίνει ότι είναι και το καλύτερο μας μοντέλο για την εξήγηση την ανεξάρτητη μεταβλητή μας

```
stepBE <- step(y4,scope=list(lower= ~1,upper= ~HT+MCH+MCV,direction="backward"))
stepBE
```

Start: **AIC=17.6**
HB ~ HT + MCH + MCV

	Df	Sum of Sq	RSS	AIC
<none>		32.313		17.595
- MCV	1	4.3876	36.701	18.142
- MCH	1	10.2210	42.534	21.092
- HT	1	12.6636	44.977	22.208

Παρατηρούμε ότι το full μοντέλο ξεκινά με AIC=17.6 και καθώς
Αφαιρει κάθε μια επεξηγηματική μεταβλητή αυτό μεγαλώνει πράγμα
που δεν το θέλουμε ,οπότε καταλήγουμε
και με αυτόν τον τρόπο όπως και ήταν αναμενόμενο ότι χρειάζονται και
οι 3 μεταβλητές μας για την βέλτιστη εξήγηση του μοντέλου μας.
(το forward elimination θα μας έβγαζε ακριβώς το ίδιο αποτέλεσμα)

ΕΥΧΑΡΙΣΤΩ