ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Συσχέτιση και Γραμμική Παλινδρόμηση

Πέτρος Κολοβός

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ

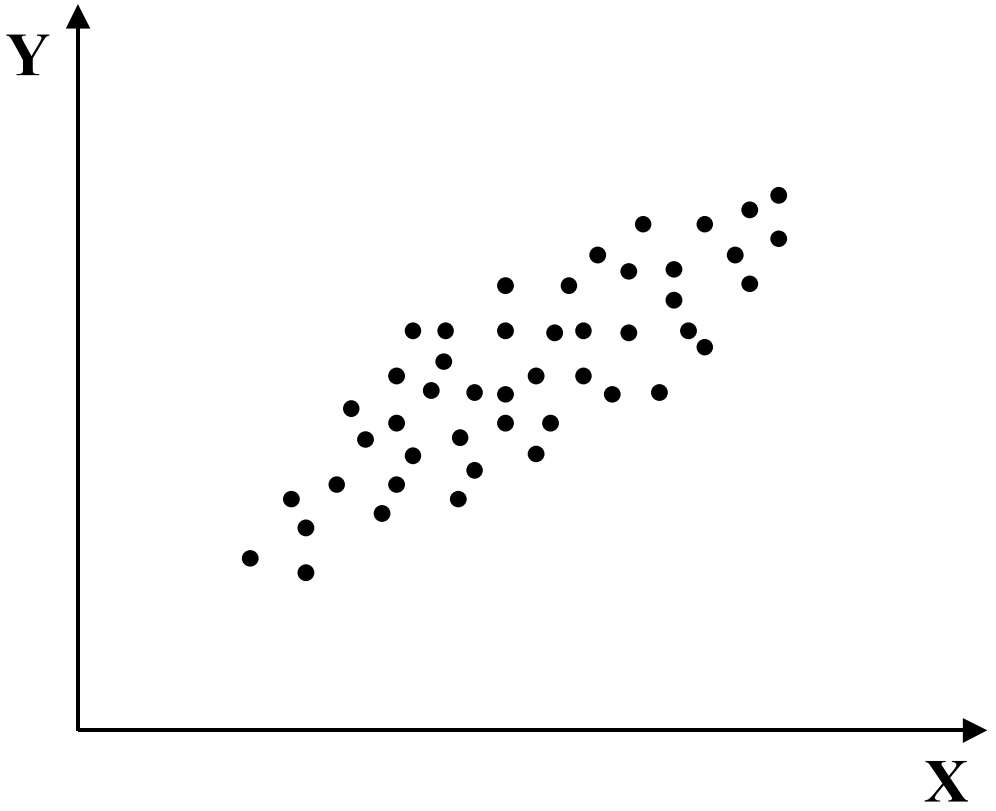
Εισαγωγή

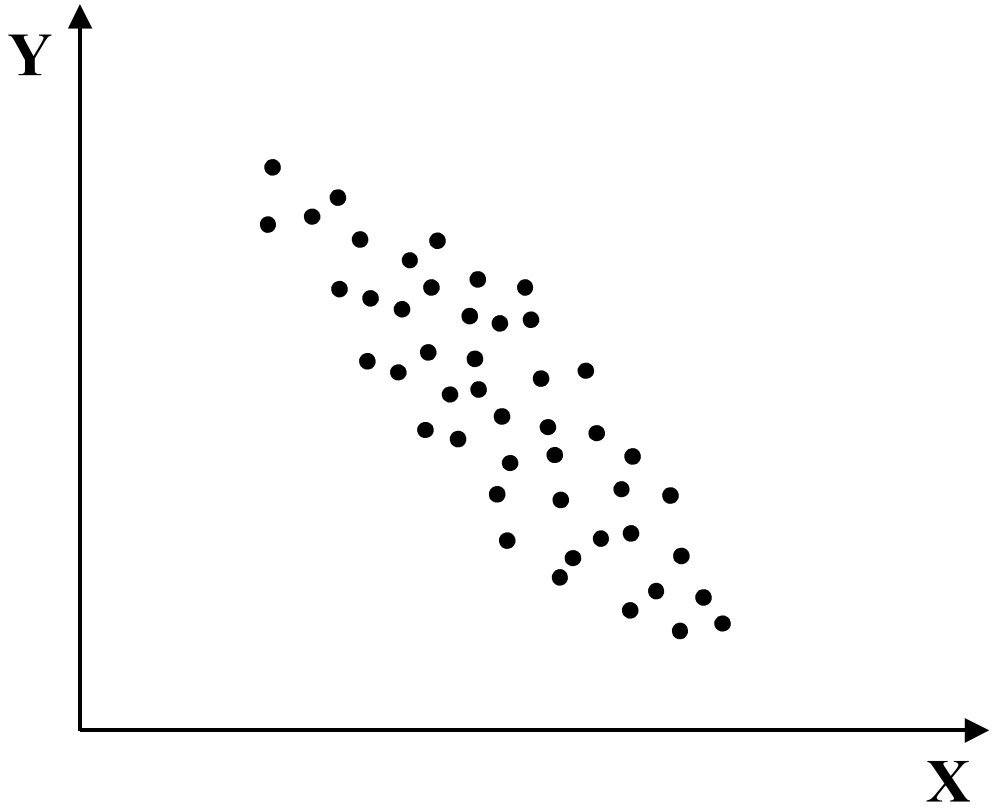
- Ειδαμε τρόπους για να συγκρίνουμε δείγματα μεταξύ τους σε επίπεδο μέσων τιμών, λόγων, εμπλουτισμών και διασποράς. Σε όλες τις περιπτώσεις που εξετάσαμε, τα δείγματα που συγκρίναμε προέρχονταν από μετρήσεις του ίδιου τύπου κάτω από διαφορετικές συνθήκες και αντιστοιχούσαν σε διαφορετικά άτομα.
- Πολλές φορές τα δεδομένα μας είναι πιο σύνθετα. Για παράδειγμα συχνά, το ίδιο άτομο έχει μετρηθεί κάτω από **διαφορετικές συνθήκες ή για διαφορετικές ιδιότητες** οπότε έχουμε αριθμητικά δεδομένα περισσότερων του ενός τύπων για κάθε στοιχείο του δείγματος. Λέμε σε αυτές τις περιπτώσεις ότι τα δεδομένα μας είναι **συζευγμένα ή ζευγαρωτά (paired)**.
- Σε πολλές περιπτώσεις επίσης έχουμε **ποσοτικά (παραμετρικά) δεδομένα διαφορετικών τύπων** εκτός από κατηγορικές μεταβλητές κι έτσι μας ενδιαφέρει **να γνωρίζουμε τις σχέσεις που τα διέπουν**. Στη βάση αυτών των ερωτημάτων βρίσκονται οι έννοιες της **συσχέτισης** και της **παλινδρόμησης** που παρότι είναι εννοιολογικά διαφορετικές έχουν κοινό χαρακτηριστικό ότι **εφαρμόζονται σε παραμετρικά, συζευγμένα δεδομένα**.

Συσχέτιση

- Η στατιστική συσχέτιση (correlation) εκφράζει τις σχέσεις μεταξύ παραμετρικών, αριθμητικών δεδομένων, μέσω των συντελεστών συσχέτισης (correlation coefficients), που παίρνουν τιμές από -1 έως 1.
- Μεταξύ των διαφορετικών ειδών συσχέτισης που μπορούμε να υπολογίσουμε, οι πιο σημαντικές και συχνά χρησιμοποιούμενες είναι η **γραμμική συσχέτιση** και οι **συσχετίσεις κατάταξης**.
- Ως τώρα εξετάσαμε μία μεταβλητή κάθε φορά
- Πολλές φορές τα δεδομένα μας είναι πιο σύνθετα
  - Το ίδιο άτομο μπορεί να έχει μετρηθεί κάτω από διαφορετικές συνθήκες ή για διαφορετικές ιδιότητες
  - Έχουμε περισσότερες μεταβλητές για κάθε στοιχείο του δείγματος
- Όταν οι τιμές δύο (ή περισσότερων) μεταβλητών δεν εμφανίζονται τυχαία αλλά με κάποια σχέση μεταξύ τους, τότε λέμε ότι οι μεταβλητές εμφανίζουν συσχέτιση

Θετική και αρνητική συσχέτιση



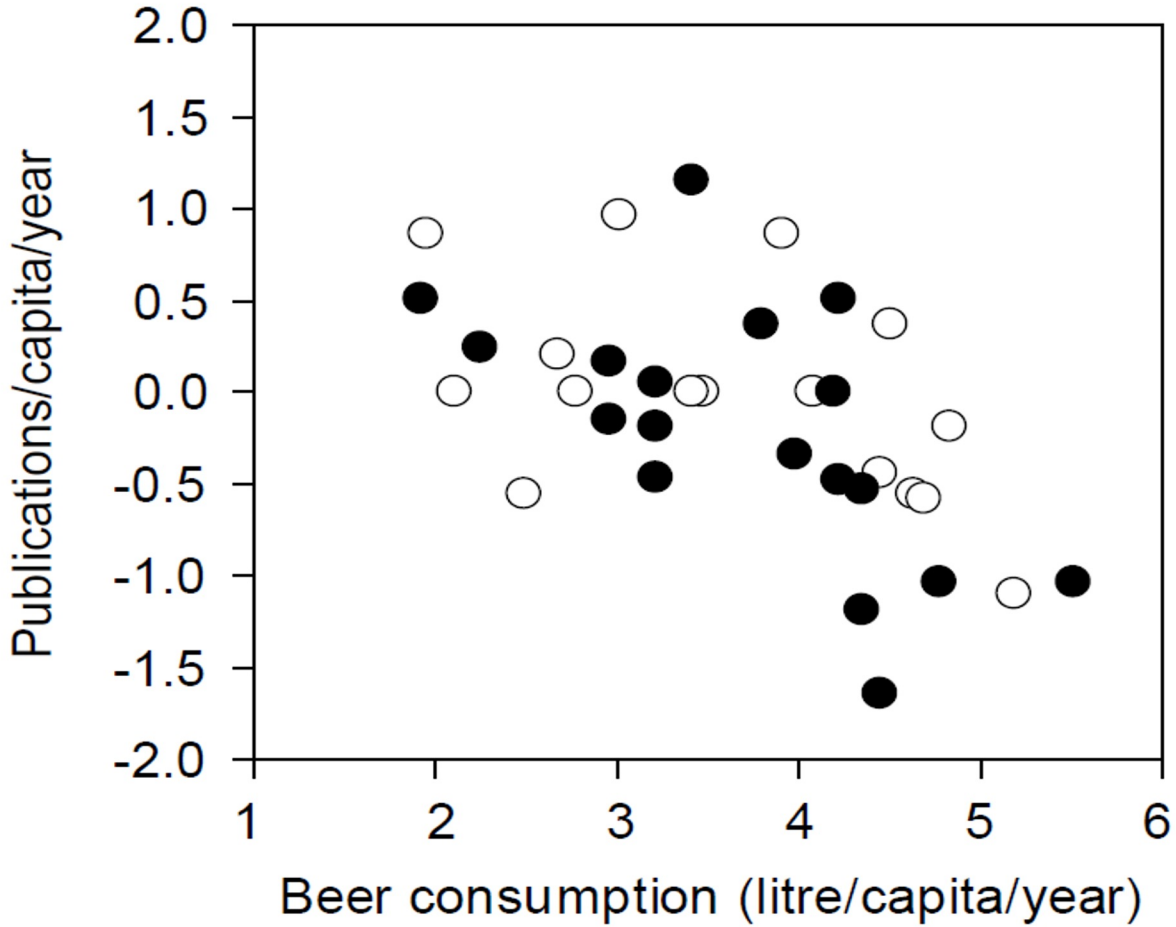


Αιτιολογημένη συσχέτιση

- Υπάρχει μια διαδικασία αιτιότητας: μια αιτία και ένα αποτέλεσμα
- Παράδειγμα: Υπάρχει μια αιτιώδης σχέση μεταξύ της θερμοκρασίας του περιβάλλοντος και της κατανάλωσης ηλεκτρικής ενέργειας
  - Η άνοδος της θερμοκρασίας κάνει τα νοικοκυριά να καταναλώνουν περισσότερη ηλεκτρική ενέργεια για λόγους κλιματισμού

Μη αιτιολογημένη συσχέτιση

- Ο ύπνος φορώντας παπούτσια είναι στενά συνδεδεμένος με το ξύπνημα με πονοκέφαλο
  - Συμπέρασμα: ο ύπνος με τα παπούτσια προκαλεί πονοκέφαλο
- Η συσχέτιση ΔΕΝ συνεπάγεται αιτιότητα
  - Μια πιο πιθανή εξήγηση είναι ότι και οι δύο επιπτώσεις προκαλούνται από έναν τρίτο παράγοντα, όπως το να πάει κανείς για ύπνο μεθυσμένος
  - Επομένως, το συμπέρασμα είναι λάθος



Oikos 117: 484–487, 2008
doi: 10.1111/j.2008.0030-1299.16551.x,

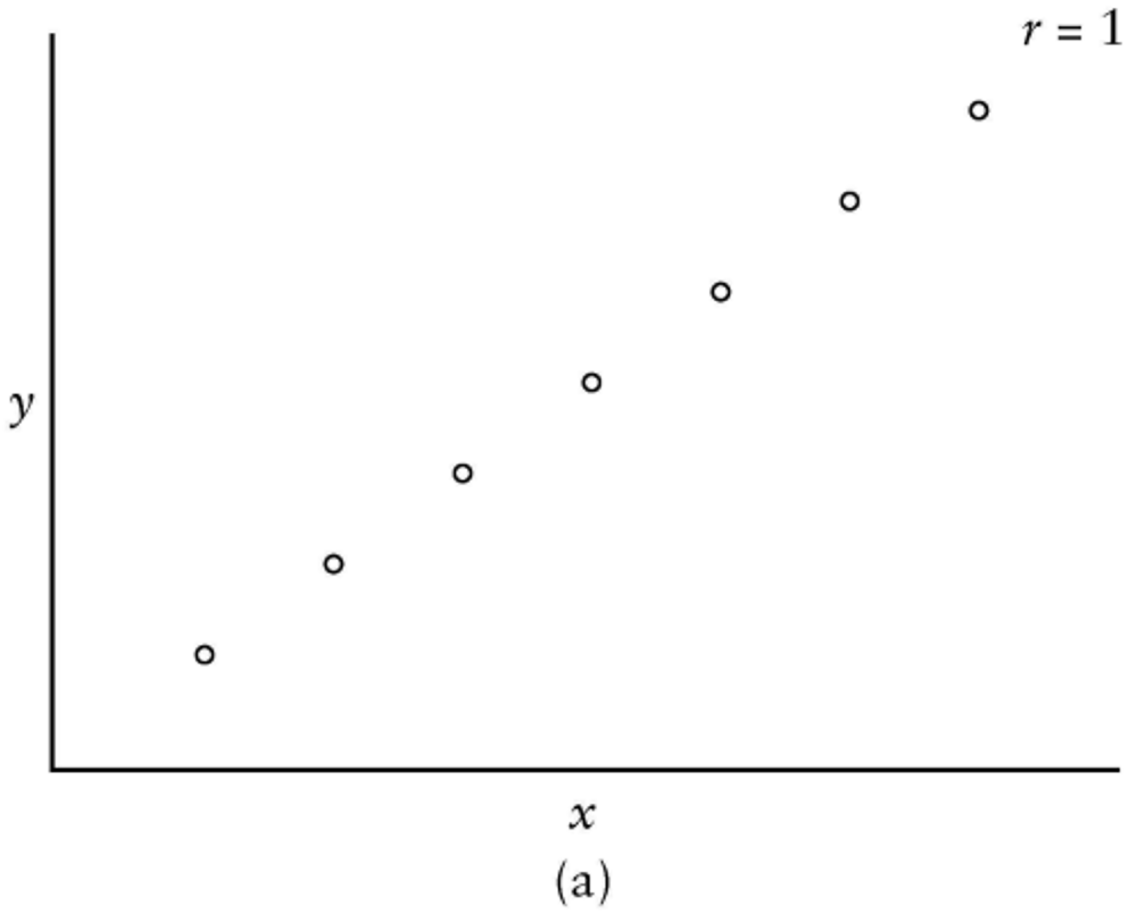
Subject Editor: Per Lundberg, Accepted 7 January 2008

A possible role of social activity to explain differences in publication output among ecologists

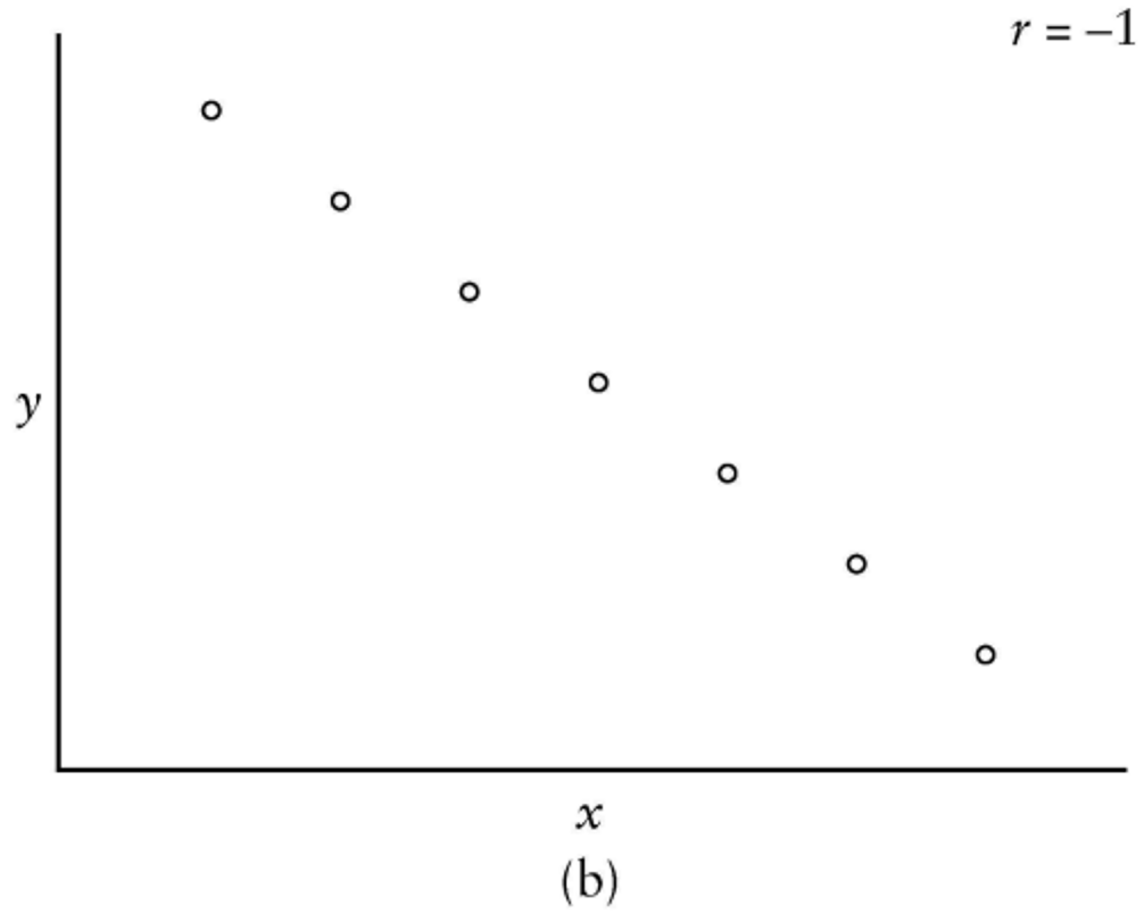
Η συσχέτιση δεν αποδεικνύει πάντα αιτιότητα

Fig. 1. Number of publications per capita per year published by Czech avian ecologists up to 2006 plotted against their beer consumption per capita per year in litres. Both data sets shown are Box-Cox transformed (thus neither the output score nor the consumption score values enable the identification of particular persons included in this research). The negative relationship between beer consumption and publication success is significant not only for the whole data set ($r_s = -0.55$, $n = 34$, $p = 0.0008$) but also for "past" (included in the first survey in 2002; ●) and "present" researchers (included in 2006; ○) analyzed separately ("past": $r_s = -0.68$, $n = 18$, $p = 0.002$; "present": $r_s = -0.52$, $n = 16$, $p = 0.04$).

Συντελεστής γραμμικής συσχέτισης r

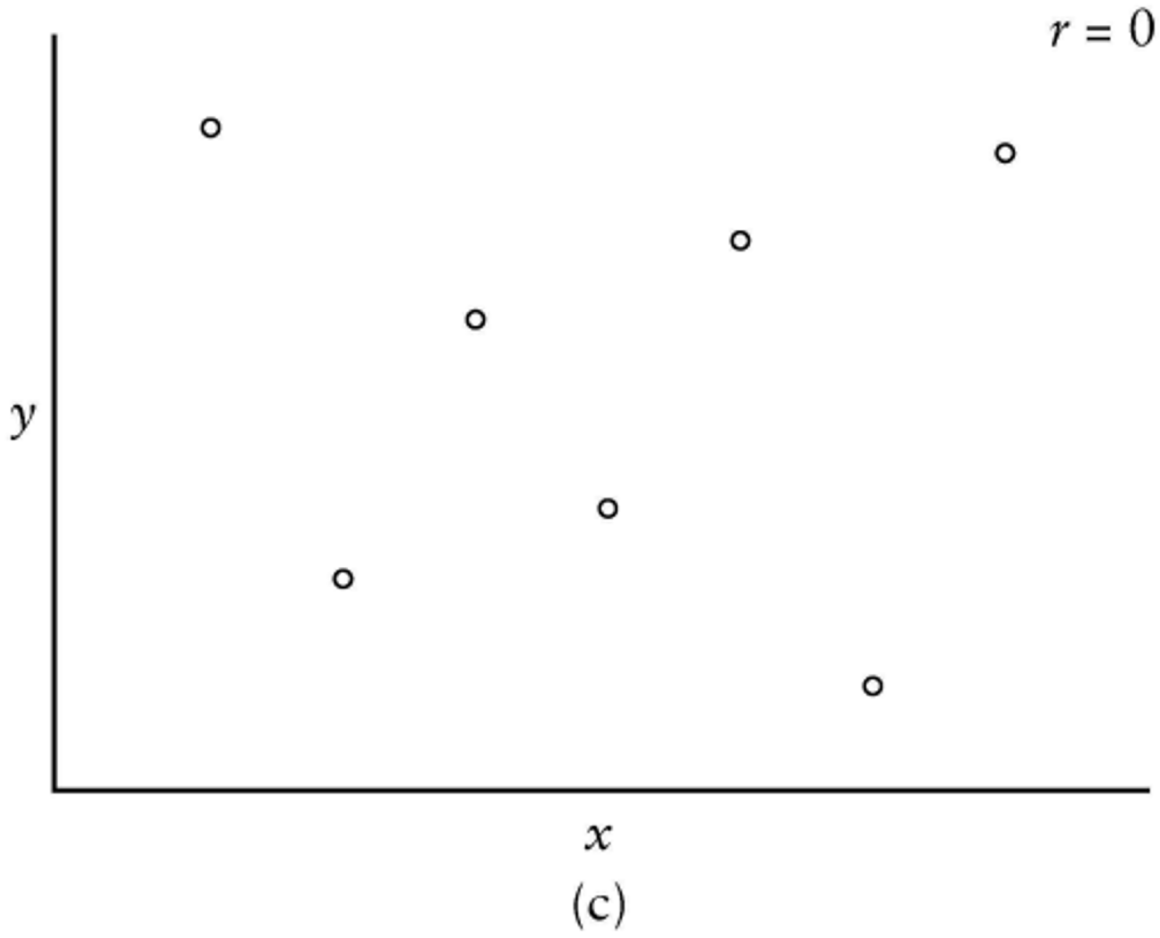


(a)

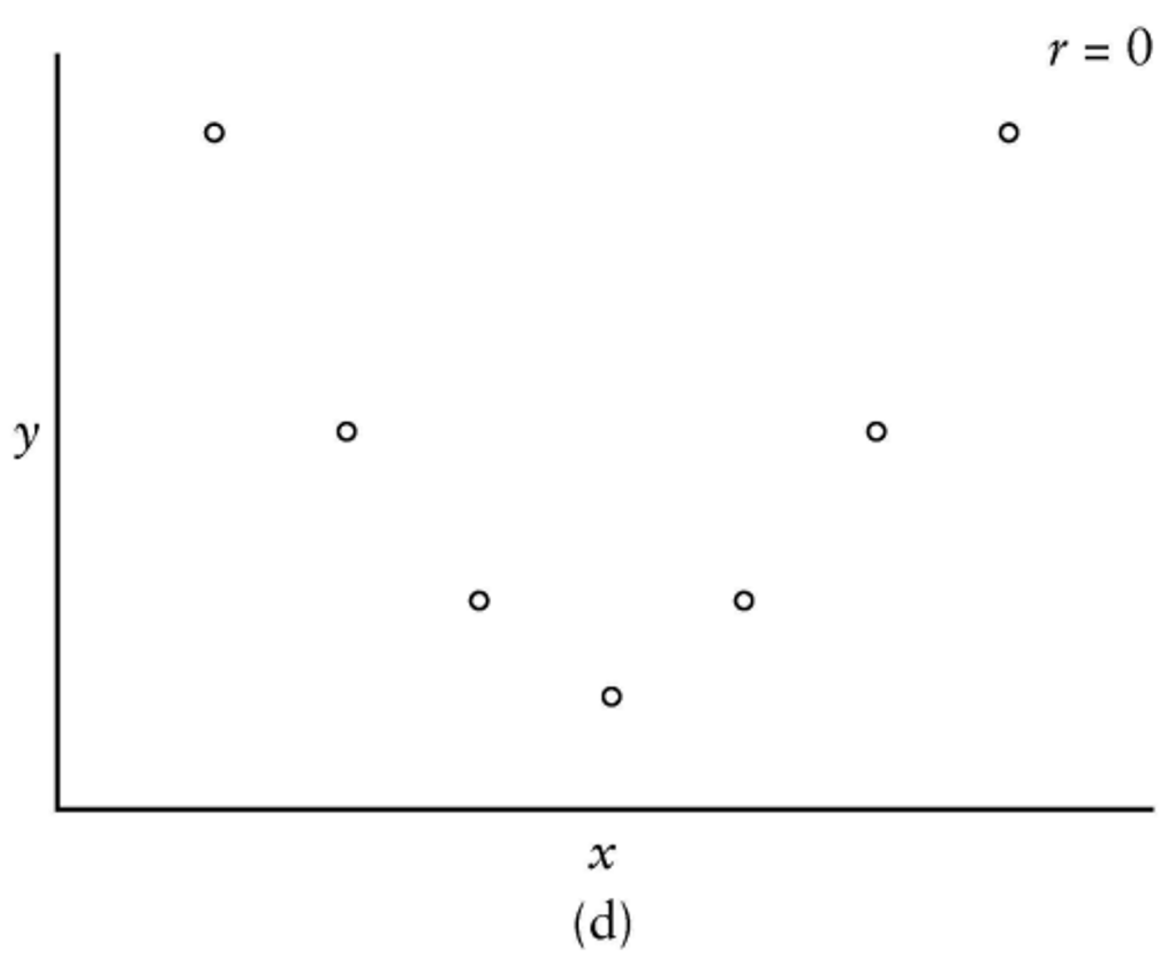


(b)

Συντελεστής γραμμικής συσχέτισης r



(c)



(d)

Γραμμική Συσχέτιση Pearson (Pearson Linear Correlation)

Θα δούμε τη γραμμική συσχέτιση μελετώντας ένα από τα ενσωματωμένα σύνολα δεδομένων της R που ονομάζεται USJudgeRatings και που περιέχει την αξιολόγηση 43 δικαστών του ανώτερου δικαστηρίου των ΗΠΑ (US Superior Court) σε 12 χαρακτηριστικά όπως η γνώση της νομοθεσίας, η προετοιμασία για την δίκη κλπ, έτσι όπως έχουν προκύψει από ερωτηματολόγια που δόθηκαν στους δικηγόρους που παραστάθηκαν σε υποθέσεις τους.

Hide

```
str(USJudgeRatings)
```

```
## 'data.frame':    43 obs. of  12 variables:
##  $ CONT: num  5.7 6.8 7.2 6.8 7.3 6.2 10.6 7 7.3 8.2 ...
##  $ INTG: num  7.9 8.9 8.1 8.8 6.4 8.8 9 5.9 8.9 7.9 ...
##  $ DMNR: num  7.7 8.8 7.8 8.5 4.3 8.7 8.9 4.9 8.9 6.7 ...
##  $ DILG: num  7.3 8.5 7.8 8.8 6.5 8.5 8.7 5.1 8.7 8.1 ...
##  $ CFMG: num  7.1 7.8 7.5 8.3 6 7.9 8.5 5.4 8.6 7.9 ...
##  $ DECI: num  7.4 8.1 7.6 8.5 6.2 8 8.5 5.9 8.5 8 ...
##  $ PREP: num  7.1 8 7.5 8.7 5.7 8.1 8.5 4.8 8.4 7.9 ...
##  $ FAMI: num  7.1 8 7.5 8.7 5.7 8 8.5 5.1 8.4 8.1 ...
##  $ ORAL: num  7.1 7.8 7.3 8.4 5.1 8 8.6 4.7 8.4 7.7 ...
##  $ WRIT: num  7 7.9 7.4 8.5 5.3 8 8.4 4.9 8.5 7.8 ...
##  $ PHYS: num  8.3 8.5 7.9 8.8 5.5 8.6 9.1 6.8 8.8 8.5 ...
##  $ RTEN: num  7.8 8.7 7.8 8.7 4.8 8.6 9 5 8.8 7.9 ...
```

Το βασικό στοιχείο είναι ότι πρόκειται για συζευγμένα ποσοτικά δεδομένα καθώς πως βλέπουμε από τις πρώτες γραμμές οι 12 μεταβλητές αντιστοιχούν στον ίδιο δικαστή (άτομο) και είναι όλες ποσοτικά δεδομένα.

Γραμμική Συσχέτιση Pearson (Pearson Linear Correlation)

Hide

```
head(USJudgeRatings)
```

```
##                CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## AARONSON,L.H.   5.7  7.9  7.7  7.3  7.1  7.4  7.1  7.1  7.1  7.0  8.3  7.8
## ALEXANDER,J.M.  6.8  8.9  8.8  8.5  7.8  8.1  8.0  8.0  7.8  7.9  8.5  8.7
## ARMENTANO,A.J.  7.2  8.1  7.8  7.8  7.5  7.6  7.5  7.5  7.3  7.4  7.9  7.8
## BERDON,R.I.     6.8  8.8  8.5  8.8  8.3  8.5  8.7  8.7  8.4  8.5  8.8  8.7
## BRACKEN,J.J.    7.3  6.4  4.3  6.5  6.0  6.2  5.7  5.7  5.1  5.3  5.5  4.8
## BURNS,E.B.      6.2  8.8  8.7  8.5  7.9  8.0  8.1  8.0  8.0  8.0  8.6  8.6
```

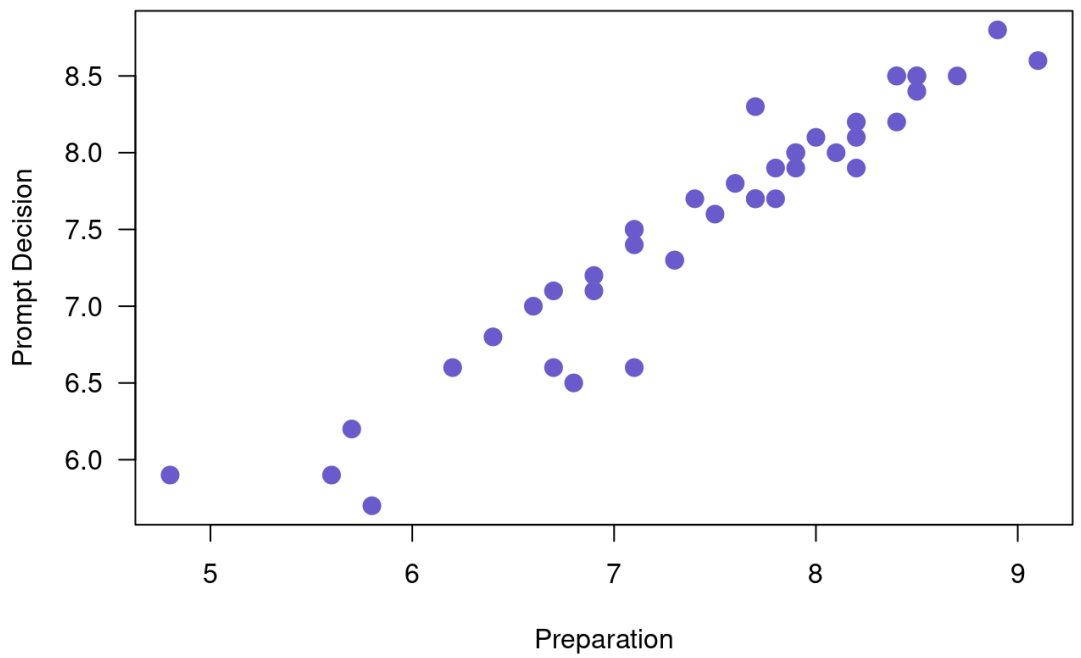
Μια από τις βασικές παραμέτρους στην εκδίκαση μιας υπόθεσης είναι ο χρόνος που χρειάζεται ο δικαστής για να εκδόσει την απόφασή του. Παρότι όλες οι υποθέσεις δεν είναι προφανώς το ίδιο εύκολες, είναι αναμενόμενο να υποθέσουμε ότι *μια σωστή προετοιμασία από πλευράς του δικαστή θα οδηγεί και σε ταχύτερη έκδοση απόφασης*.

Γραφική διερεύνηση της συσχέτισης

ΥΠΟΘΕΣΗ: Έστω ότι μας ενδιαφέρει να δούμε αν υπάρχει μια συσχέτιση μεταξύ του χρόνου έκδοση της απόφασης (που δίνεται από την μεταβλητή $DECI$) και της προετοιμασίας ($PREP$). Μπορούμε αρχικά να δοκιμάσουμε να αναπαραστήσουμε τα δεδομένα γραφικά με ένα διάγραμμα σκέδασης:

Hide

```
plot(USJudgeRatings$DECI~USJudgeRatings$PREP, type="p", pch=19, col="slateblue", las=1, ylab="Prompt De
cision", xlab="Preparation", cex=1.4)
```



Από το διάγραμμα βλέπουμε ότι μια γραμμική σχέση είναι κάτι παραπάνω από πιθανή. Όσο αυξάνονται τα score που αποτιμούν την προετοιμασία (οριζόντιος άξονας) τόσο αυξάνεται και η ταχύτητα έκδοσης απόφασης (κάθετος άξονας).

Υπολογισμός γραμμικής συσχέτισης

Μπορούμε τώρα να υπολογίσουμε τον συντελεστή Pearson με μια απλή κλήση της συναρτησης $cor()$

Hide

```
cor(USJudgeRatings$DECI, USJudgeRatings$PREP, method="pearson")
```

```
## [1] 0.9570883
```

Βλέπουμε ότι η τιμή είναι πολύ υψηλή και θετική, που σημαίνει ότι υπάρχει μια έντονη θετική συσχέτιση μεταξύ των δύο ποσοτήτων όπως θα περιμέναμε.

Στατιστική εκτίμηση της συσχέτισης

Αν θέλουμε να αξιολογήσουμε την στατιστική σημασία αυτής της συσχέτισης μπορούμε να εφαρμόσουμε έναν έλεγχο συσχέτισης με την συνάρτηση $cor.test()$:

Hide

```
cor.test(USJudgeRatings$DECI, USJudgeRatings$PREP, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  USJudgeRatings$DECI and USJudgeRatings$PREP
## t = 21.147, df = 41, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9216892 0.9766800
## sample estimates:
##       cor
## 0.9570883
```

απ' όπου προκύπτει ότι η μηδενική υπόθεση για συσχέτιση ίση με το 0 απορρίπτεται στη βάση τόσο της πολύ τιμής p-value όσο και του διαστήματος εμπιστοσύνης.

Συσχέτιση και ελλιπείς τιμές

Ας δούμε άλλο ένα παράδειγμα από ένα άλλο σύνολο δεδομένων το *airquality* που περιέχει μετρήσεις μετεωρολογικών ενδείξεων για την πόλη της Νέας Υόρκης μεταξύ του Μαΐου και Σεπτεμβρίου του 1973.

Hide

```
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

Ας δοκιμάσουμε να δούμε την έκτασης της συσχετίσης των επιπέδων του όζοντος και της έντασης της ηλιακής ακτινοβολίας:

Hide

```
cor(airquality$Ozone, airquality$Solar.R)
```

```
## [1] NA
```

Αν προσέξετε καλύτερα θα δείτε ότι το αρχικό σύνολο δεδομένων έχει τιμές “NA” δηλαδή ελλιπείς μετρήσεις για κάποιες ημέρες.

Η *cor*() μπορεί να μη λάβει υπ’ όψιν τις ελλιπείς τιμές με την χρήση της παραμέτρου *use* =:

Hide

```
cor(airquality$Ozone, airquality$Solar.R, use="complete.obs")
```

```
## [1] 0.3483417
```

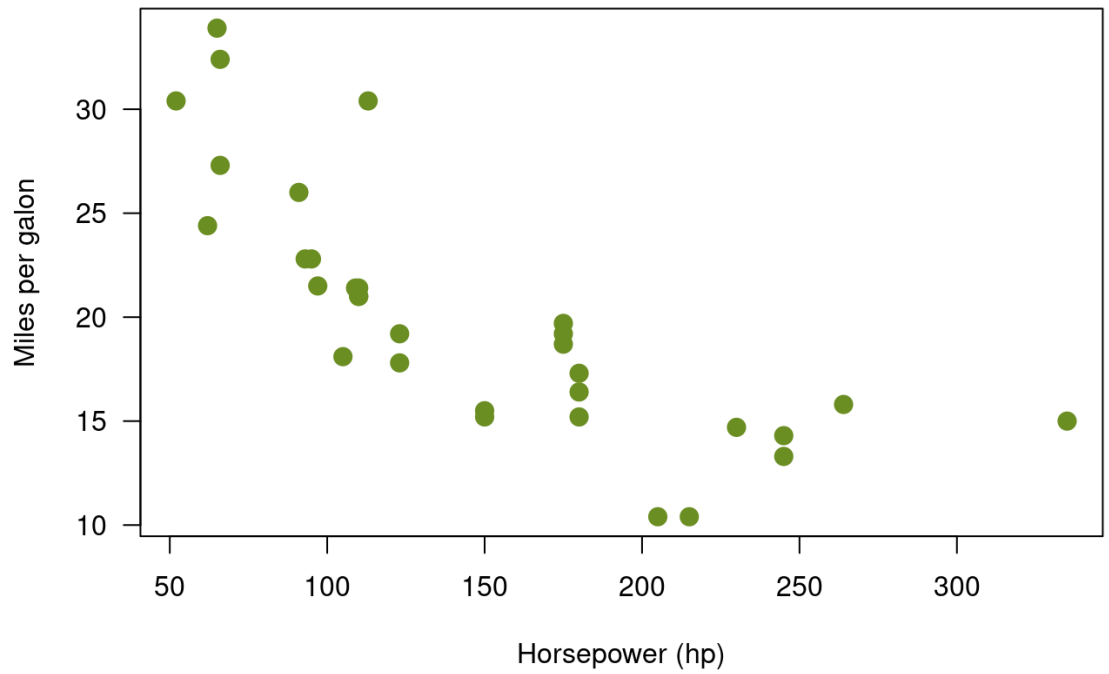
Συσχετίσεις κατάταξης (Rank Correlations)

Υπάρχουν περιπτώσεις που το κριτήριο της γραμμικότητας δεν ισχύει, η σχέση δηλαδή μεταξύ των δύο μεταβλητών δεν μπορεί να αναπαρασταθεί γραφικά με μια απλή ευθεία γραμμή.

Δείτε για παράδειγμα την αρνητική σχέση μεταξύ κατανάλωσης καυσίμου (miles per galon, mpg) και ιπποδύναμης (horsepower, hp) στο σύνολο δεδομένων $mtcars$.

Hide

```
plot(mtcars$mpg~mtcars$hp, type="p", pch=19, col="olivedrab", las=1, ylab="Miles per galon", xlab="Horsepower (hp)", cex=1.4)
```



Βλέπουμε ότι η σχέση είναι αντίστροφη, δηλαδή όσο αυξάνεται η ιπποδύναμη τόσο μειώνεται ο αριθμός των μιλίων ανα γαλόνι (και άρα αυξάνεται η κατανάλωση). Αν υπολογίσουμε την συσχέτιση με βάση την γραμμική σχέση:

Συσχετίσεις κατάταξης (Rank Correlations)

Βλέπουμε ότι η σχέση είναι αντίστροφη, δηλαδή όσο αυξάνεται η ιπποδύναμη τόσο μειώνεται ο αριθμός των μιλίων ανα γαλόνι (και άρα αυξάνεται η κατανάλωση). Αν υπολογίσουμε την συσχέτιση με βάση την γραμμική σχέση:

Hide

```
cor.test(mtcars$mpg, mtcars$hp, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$mpg and mtcars$hp
## t = -6.7424, df = 30, p-value = 1.788e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8852686 -0.5860994
## sample estimates:
##       cor
## -0.7761684
```

παίρνουμε μια ισχυρά αρνητική συσχέτιση που είναι στατιστικά σημαντική.

Κοιτώντας και πάλι το διάγραμμα ωστόσο, *είναι δύσκολο να φανταστούμε μια ευθεία γραμμή γύρω από την οποία να κατανέμονται τα σημεία*. Στην περίπτωση αυτή είναι προτιμότερο να εφαρμόσουμε μια συσχέτιση κατάταξης.

Συσχετίσεις κατάταξης (Rank Correlations)

Ο συντελεστής rho του Spearman είναι ένα μέτρο συσχέτισης κατάταξης. Μπορούμε να τον υπολογίσουμε επιλέγοντας την αντίστοιχη μέθοδο στη συνάρτηση *cor*() ή *cor. test*():

Hide

```
cor.test(mtcars$mpg, mtcars$hp, method="spearman")
```

```
## Warning in cor.test.default(mtcars$mpg, mtcars$hp, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  mtcars$mpg and mtcars$hp
## S = 10337, p-value = 5.086e-12
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.8946646
```

Γενικά εφαρμόζουμε συσχέτιση κατάταξης όταν δεν έχουμε γραμμικές σχέσεις αλλά και όταν έχουμε ακραίες τιμές (outliers) καθώς ο συντελεστής Pearson είναι πολύ πιο “ευαίσθητος” σε αυτές σε σχέση με τους συντελεστές κατάταξης.

Συσχέτιση κατάστασης κατά Kendall

Η συσχέτιση κατά Spearman λειτουργεί ικανοποιητικά σε περιπτώσεις που οι μεταβλητές μας είναι κανονικά κατανεμημένες. Στην αντίθετη περίπτωση χρησιμοποιούμε και πάλι συσχέτιση κατάταξης όμως αυτή την φορά χρησιμοποιούμε τον συντελεστή tau του Kendall.

Ο συντελεστής Kendall είναι πιο αυστηρός από τον Spearman κι έτσι σχεδόν πάντοτε δίνει χαμηλότερες σε απόλυτη τιμή εκτιμήσεις. Το παράδειγμά μας δεν αποτελεί εξαίρεση:

Hide

```
cor.test(mtcars$mpg, mtcars$hp, method="kendall")
```

```
## Warning in cor.test.default(mtcars$mpg, mtcars$hp, method = "kendall"): Cannot
## compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  mtcars$mpg and mtcars$hp
## z = -5.871, p-value = 4.332e-09
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.7428125
```


Γενικά κριτήρια για την εφαρμογή συσχετίσεων

Σε γενικές γραμμές, σε ό,τι αφορά τις συσχετίσεις οι κανόνες που πρέπει να θυμόμαστε είναι:

- Χρησιμοποιούμε την γραμμική συσχέτιση Pearson σε γραμμικές σχέσεις με κανονικά κατανεμημένες μεταβλητές χωρίς ακραίες τιμές.
- Χρησιμοποιούμε την συσχέτιση Spearman σε μη-γραμμικές κανονικά κατανεμημένες τιμές.
- Χρησιμοποιούμε την συσχέτιση Kendall σε μη-γραμμικές, μη-κανονικά κατανεμημένες τιμές.

Συνδιακύμανση και μερική συσχέτιση (covariance)

Από πρακτικής άποψης η συσχέτιση μας λέει σε ποιο βαθμό δύο μεταβλητές μεταβάλλονται με αντίστοιχο τρόπο, δηλαδή συνδιακυμαίνονται. Στη στατιστική **η συσχέτιση είναι μια κανονικοποιημένη μορφή της συνδιακύμανσης** που παίρνει τιμές μεταξύ -1 και 1.

Η R εχει μια απλή συνάρτηση για την συνδιακύμανση, που ονομάζεται $cov()$ και που λειτουργεί ακριβώς όπως και η $cor()$:

Hide

```
cov(mtcars$mpg, mtcars$hp)
```

```
## [1] -320.7321
```

Όπως βλέπουμε από το παράδειγμα η συνδιακύμανση παίρνει τιμές που μπορεί να είναι πολύ μεγάλες ανάλογα με τα εύρη των τιμών των μεταβλητών που εξετάζουμε.

Από πρακτικής άποψης η συνδιακύμανση είναι πολύ λιγότερο χρήσιμη σε σχέση με την συσχέτιση ακριβώς λόγω αυτής της εξάρτησής της από την κλίμακα των μεταβλητών. Ωστόσο την χρησιμοποιούμε σε περιπτώσεις που θέλουμε να υπολογίσουμε "μερικές συσχετίσεις" (partial correlations) καθώς τότε είναι σημαντικό να εκτιμήσουμε τον τρόπο με τον οποίο συνδιακυμαίνονται περισσότερες από μία μεταβλητές.

Τι είναι η μερική συσχέτιση;

Φανταστείτε τρεις μεταβλητές (a, b, c) που έχουμε μετρήσει για τα ίδια στοιχεία. Μπορούμε να υπολογίσουμε την συσχέτισή τους ανά δυο, ωστόσο τίποτα δε μας λέει ότι οι τιμές που προκύπτουν για κάθε ζεύγος δεν εξαρτώνται εν μέρει και από την τρίτη μεταβλητή.

Αν επιστρέψουμε στο παράδειγμα της κατανάλωσης σε σχέση με την ιπποδύναμη, η αρνητική συσχέτιση μπορεί να οφείλεται εν μέρει σε μια άλλη τάση που έχουν τα αυτοκίνητα με πολλούς ίππους π.χ. να είναι πιο βαριά λόγω της μεγαλύτερης μηχανής τους.

Μπορούμε να υπολογίσουμε όλες τις συσχετίσεις μεταξύ των τριών αυτών μεταβλητών δίνοντας στην $cor()$ έναν πίνακα που τις περιέχει και τις τρεις:

Hide

```
cor(mtcars[,c(1,4,6)])
```

```
##            mpg         hp         wt
## mpg  1.0000000 -0.7761684 -0.8676594
## hp  -0.7761684  1.0000000  0.6587479
## wt  -0.8676594  0.6587479  1.0000000
```

Εδώ βλέπουμε ότι η κατανάλωση είναι αρνητικά συσχετισμένη **και** με την ιπποδύναμη αλλά **και** με το βάρος.

Το ερώτημα είναι: **Σε ποιο βαθμό η συσχέτιση με την πρώτη εξαρτάται από το δεύτερο;**

Υπολογισμός μερικών συσχετίσεων

Μπορούμε να υπολογίσουμε την μερική συσχέτιση κατανάλωσης-ιπποδύναμης ελέγχοντας για το βάρος με την χρήση της συνάρτησης $pcor()$ από το πακέτο ggm ως εξής:

Hide

```
library(ggm)
pcor(c(1,4,6), cov(mtcars))
```

```
## [1] -0.5469926
```

Η σύνταξη της $pcor()$ δέχεται ως πρώτο όρισμα έναν πίνακα που περιέχει τις δύο προς σύγκριση μεταβλητές πρώτες και τις μεταβλητές ελέγχου στη συνέχεια, ακολουθούμενες από τον πίνακα συνδιακύμανσης όλων των δεδομένων.

Εδώ οι μεταβλητές που συγκρίνουμε είναι η πρώτη και τέταρτη στήλη του $mtcars$ (mpg και hp αντίστοιχα) και ο έλεγχος γίνεται ως προς την έκτη (wt).

Βλέπουμε ότι από την αρχική συσχέτιση (-0.77) ένα μέρος (|0.77-0.54|=0.23) οφείλεται στη σχέση που υπάρχει με την τρίτη μεταβλητή (wt) .

Αντίστοιχα μπορούμε να ελέγξουμε την μερική συσχέτιση για περισσότερες από μία μεταβλητές. Για παράδειγμα αν θέλουμε να ελέγξουμε όχι μόνο για το βάρος αλλά και για άλλα χαρακτηριστικά της μηχανής όπως ο αριθμός των κυλίνδρων (cyl) και ο όγκος τους ($disp$):

Hide

```
pcor(c(1,4,6,2,3), cov(mtcars))
```

```
## [1] -0.3094339
```

Κάτι που σημαίνει ότι από την αρχική συσχέτιση μεταξύ κατανάλωσης και ιπποδύναμης, περισσότερη από την μισή μπορεί να αποδοθεί σε χαρακτηριστικά της μηχανής.

Παλινδρόμηση (Regression)

- *Η συσχέτιση μας λέει κατά πόσο υπάρχει μια σχέση μεταξύ δύο αριθμητικών μεταβλητών*
  - η **παλινδρόμηση** είναι μια δέσμη στατιστικών εργαλείων για να αποτιμήσουμε τη σχέση αυτή ποσοτικά
- *Π.χ. η κατανάλωση βενζίνης έχει ισχυρή συσχέτιση με το βάρος του αυτοκινήτου*
  - Πώς όμως θα μπορούσαμε να εκτιμήσουμε την κατανάλωση που θα έχει ένα αυτοκίνητο με δεδομένο βάρος;
  - Θα χρειαστούμε μια αριθμητική σχέση, δηλαδή μια εξίσωση που θα περιέχει το βάρος και την κατανάλωση ως μεταβλητές

Παλινδρόμηση (Regression)

Αν η συσχέτιση μας λέει κατά πόσο υπάρχει μια σχέση μεταξύ δύο αριθμητικών μεταβλητών, η παλινδρόμηση είναι μια δέσμη στατιστικών εργαλείων για να τις αποτιμήσουμε ποσοτικά.

Στο απλό παράδειγμα που είδαμε παραπάνω, υπολογίσαμε ότι η κατανάλωση βενζίνης έχει ισχυρή συσχέτιση με το βάρος του αυτοκινήτου. Πώς όμως θα μπορούσαμε να εκτιμήσουμε την κατανάλωση που θα έχει ένα αυτοκίνητο με δεδομένο βάρος;

Τι χρειαζόμαστε: Θα χρειαστούμε μια αριθμητική σχέση, δηλαδή **μια εξίσωση που θα περιέχει το βάρος και την κατανάλωση ως μεταβλητές**.

Πώς ορίζουμε το μοντέλο: Στην προκειμένη περίπτωση λέμε ότι το βάρος είναι η ανεξάρτητη, (ή επεξηγηματική, predictor variable) μεταβλητή και η κατανάλωση η εξαρτημένη μεταβλητή (ή μεταβλητή απόκρισης, response variable).

Ιδανικά δηλαδή θα επιθυμούσαμε να γνωρίζουμε μια σχέση του γενικού τύπου:

> Κατανάλωση = f(βάρος)

Ανάλυση Παλινδρόμησης

Το αντικείμενο της ανάλυσης παλινδρόμησης είναι η εξαγωγή των όρων της $f()$ στην σχέση:

> Μεταβλητή Στόχος = f(επεξηγηματικές μεταβλητές)

Τα είδη της παλινδρόμησης είναι:
α. απλή (μία επεξηγηματική μεταβλητή)
β. πολλαπλή (πολλές επεξηγηματικές μεταβλητές)

Ανάλογα με την μορφή της συνάρτησης μπορούμε να έχουμε:
α. γραμμική,
β. πολυωνυμική,
γ. λογιστική (logistic),
δ. Poisson ή άλλες.

Παλινδρόμηση (Regression)

Απλή Γραμμική Παλινδρόμηση

Θα επικεντρωθούμε στο πιο απλό είδος της απλής γραμμικής παλινδρόμησης για να δούμε τα βασικά στοιχεία της.

Η απλή γραμμική παλινδρόμηση (simple linear regression) είναι μια οικογένεια στατιστικών μοντέλων στα οποία μια μεταβλητή απόκρισης y εκφράζεται ως γραμμική σχέση μιας επεξηγηματικής μεταβλητής x:

$$y = \alpha x + \beta$$

και της οποίας σκοπός είναι να εκτιμήσει τις τιμές των α και β οι οποίες αντιστοιχούν στον συντελεστή του x και τον σταθερό όρο (intercept).

Η R διαθέτει μια σειρά από πακέτα και συναρτήσεις για την εκτίμηση αυτών των παραμέτρων με κυριότερη την $lm()$ την οποία θα χρησιμοποιήσουμε στη συνέχεια του κεφαλαίου για να δούμε:
1. πώς υπολογίζουμε τις παραμέτρους του μοντέλου,
2. πώς αξιολογούμε την σημασία τους και
3. πώς διακρίνουμε μεταξύ χαρακτηριστικών τιμών των δεδομένων μας.

Παλινδρόμηση με την συνάρτηση $lm()$

Θα ξεκινήσουμε με την ανάλυση παλινδρόμησης σε ένα σύνολο δεδομένων δημογραφικών στοιχείων από τις ΗΠΑ που περιέχεται στο πλαίσιο δεδομένων $state.x77$ της βασικής έκδοσης της R.

Στη συνέχεια θα δούμε πώς μπορούμε να εκτιμήσουμε την επίδραση των διαφόρων δημογραφικών στοιχείων σε δεδομένα όπως το μέσο εισόδημα και η εγκληματικότητα μέσω της εφαρμογής της απλής και της πολλαπλής γραμμικής παλινδρόμησης.

Απλη Παλινδρόμηση

Πριν ξεκινήσουμε μια ανάλυση, απαραίτητο βήμα είναι η προετοιμασία των δεδομένων στην κατάλληλη μορφή. Για την εφαρμογή της γραμμικής παλινδρόμησης μέσω της συνάρτησης $lm()$ θα χρειαστούμε ένα πλαίσιο δεδομένων (dataframe) κι έτσι θα δημιουργήσουμε ένα από την αρχική μεταβλητή $state.x77$ που είναι της μορφής matrix.

Hide

```
class(state.x77)
```

```
## [1] "matrix" "array"
```

Hide

```
state.data<-as.data.frame(state.x77)
class(state.data)
```

```
## [1] "data.frame"
```

Hide

```
str(state.data)
```

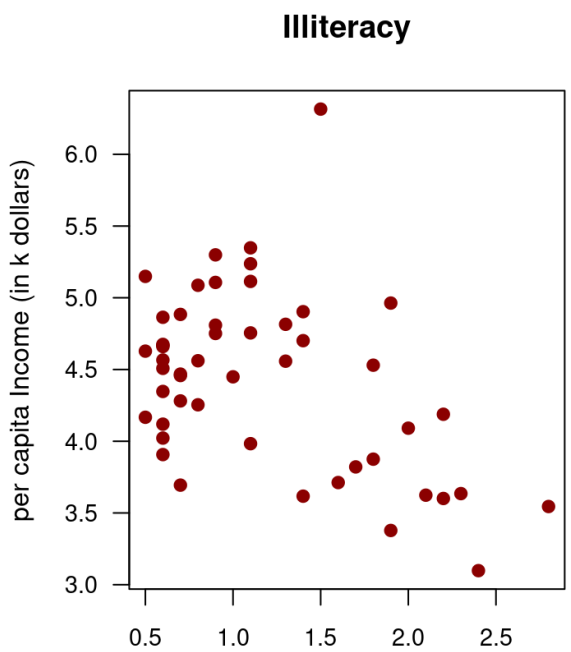
```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

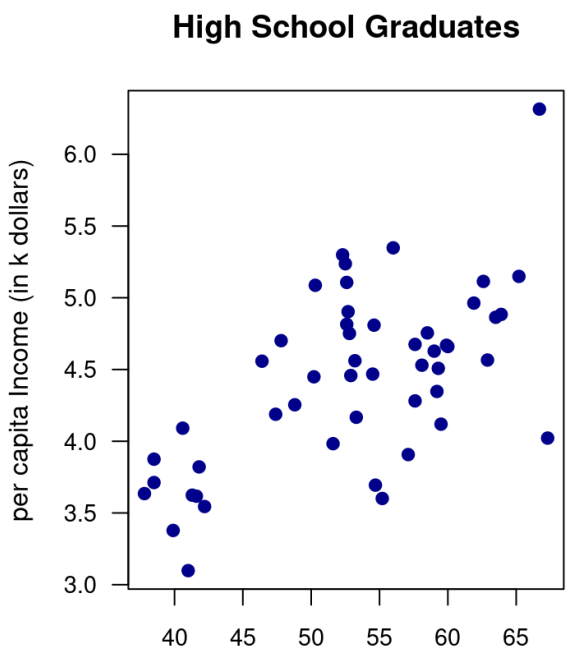
Απλη Παλινδρόμηση

Ένα επόμενο είναι η γραφική αναπαράσταση των δεδομένων που πρόκειται να αναλύσουμε. Στην συγκεκριμένη περίπτωση θέλουμε να εξετάσουμε την πιθανή εξάρτηση του μέσου εισοδήματος αρχικά με το βαθμό αναλφαβητισμού και στην συνέχεια με το ποσοστό των αποφοίτων μέσης εκπαίδευσης. Οι δύο γραφικές παραστάσεις είναι:

Hide

```
par(mfrow=c(1,2))
plot(state.data$Illiteracy, state.data$Income/1000, type="p", pch=19, col="dark red", ylab="per capita
Income (in k dollars)", xlab="Illiterate Population (%)", cex.lab=1, cex.axis=0.9, las=1, main="Illiter
acy", cex.main=1.2)
plot(state.data$`HS Grad`, state.data$Income/1000, type="p", pch=19, col="dark blue", ylab="per capita
Income (in k dollars)", xlab="HS Grads in Population (%)", cex.lab=1, cex.axis=0.9, las=1, main="High S
chool Graduates", cex.main=1.2)
```





Απλη Παλινδρόμηση

Ας εφαρμόσουμε την συνάρτηση $lm()$ για την πρώτη περίπτωση.

Στο πρώτο αυτό μοντέλο γραμμικής παλινδρόμησης έχουμε διαιρέσει με 1000 την μεταβλητή-στόχο $Income$ ώστε να εκφράζεται σε χιλιάδες δολάρια.

Hide

```
lmfit.illit<-lm(Income/1000 ~ Illiteracy, data=state.data)
summary(lmfit.illit)
```

```
##
## Call:
## lm(formula = Income/1000 ~ Illiteracy, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94889 -0.37620 -0.04977  0.34700  2.02460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9513     0.1723  28.739  < 2e-16 ***
## Illiteracy   -0.4406     0.1309  -3.367  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5584 on 48 degrees of freedom
## Multiple R-squared:  0.191,  Adjusted R-squared:  0.1742
## F-statistic: 11.34 on 1 and 48 DF,  p-value: 0.001505
```

- Αρχικά εφαρμόσαμε την $lm()$ με σύνταξη ανάλογη με αυτήν της ANOVA στο προηγούμενο κεφάλαιο (θυμηθείτε ότι και η ANOVA ανήκει στην γενικότερη κατηγορία των γραμμικών μοντέλων).
- Η μεταβλητή απόκρισης ($Income$) τοποθετείται αριστερά από την περισπωμένη (~) και η επεξηγηματική μεταβλητή (ή οι επεξεγηματικές μεταβλητές στην πολλαπλή παλινδρόμηση) δεξιά.
- Το πλαίσιο δεδομένων από το οποίο αντλούμε τις μεταβλητές ακολουθεί μετά την παράμετρο ($data =$).
- Την εκτέλεση του μοντέλου αποδίδουμε σε μια νέα μεταβλητή που ονομάζουμε $lmfit.illit$ και στη συνέχεια
- εφαρμόζουμε σε αυτό το αντικείμενο την συνάρτηση $summary()$ που επίσης είδαμε και στο κεφάλαιο της ANOVA.

Ανάλυση αποτελεσμάτων της lm()

```
##
## Call:
## lm(formula = Income/1000 ~ Illiteracy, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94889 -0.37620 -0.04977  0.34700  2.02460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9513     0.1723  28.739  < 2e-16 ***
## Illiteracy   -0.4406     0.1309  -3.367  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5584 on 48 degrees of freedom
## Multiple R-squared:  0.191,  Adjusted R-squared:  0.1742
## F-statistic: 11.34 on 1 and 48 DF,  p-value: 0.001505
```

Ας εξετάσουμε πιο προσεκτικά τι περιέχει το αποτέλεσμα της *summary*():

1. Call. Περιέχει τον τύπο (formula) της παλινδρόμησης που εκτελέσαμε. Προσέξτε ότι αντί για το πιο απλό $Income \sim Illiteracy$, έχουμε εκτελέσει ένα μοντέλο με μεταβλητή απόκρισης το $Income/1000$ για να υπολογίσουμε το μέσο εισόδημα σε χιλιάδες δολάρια.
2. Residuals. Οι αποκλίσεις (residuals) είναι οι διαφορές των τιμών της μεταβλητής απόκρισης από αυτές που εκτιμώνται από το μοντέλο και κατά συνέπεια αντανακλούν το βαθμό της αποτελεσματικότητας του μοντέλου.
3. Coefficients. Είναι το βασικό αποτέλεσμα του μοντέλου και αντιστοιχεί στους συντελεστές της εξίσωσης ευθείας που έχει προσδιοριστεί με βάση τα δεδομένα για την σχέση $y = \alpha x + \beta$. Στην περίπτωση που εξετάζουμε οι τιμές είναι $y = 4.9513 - 0.4406x$ οπως προκύπτει από την στήλη estimate. Οι τιμές αυτές ελέγχονται και από τις αντίστοιχες τιμές $p - value$ στην τελευταία στήλη, που είναι ενδεικτικές για την σημασία του κάθε παράγοντα.
4. Residual standard error. Είναι το τυπικό σφάλμα αποκλίσεων και αποτελεί μια ενδεικτική τιμή του πόσο καλά περιγράφει τα δεδομένα μας το μοντέλο. Ένα residual standard error=0.5584 σημαίνει ότι το σφάλμα στις προβλέψεις του μοντέλου σε σχέση με τα πραγματικά δεδομένα είναι περίπου 0.5584*1000=558.4 δολάρια.
5. Multiple and Adjusted R-squared. Οι τιμές αυτές είναι συντελεστές συσχέτισης. Η τιμή Multiple R-squared σημαίνει ότι το τετράγωνο του συντελεστή γραμμικής συσχέτισης κατά Pearson (βλ. παραπάνω) μεταξύ Income/Illiteracy είναι 0.191, κατά συνέπεια ο συντελεστής συσχέτισης είναι ίσος με $\sqrt{(0.191)}$. Πράγματι:

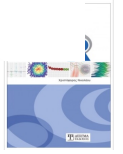
Hide

```
cor(state.data$Income, state.data$Illiteracy)**2
```

```
## [1] 0.1910347
```

Σημείωση: Ο συντελεστής Adjusted R-squared είναι μια ένδειξη του αν το μοντέλο δουλεύει καλά ή όχι. Αν ο adjusted R-squared είναι μεγαλύτερος από τον multiple R-squared τότε αυτό σημαίνει ότι το μοντέλο προβλέπει καλύτερα απ' όσο θα περιμέναμε αν κάναμε μια τυχαία εκτίμηση με μια επεξηγηματική μεταβλητή. Αν όχι, τότε αυτό σημαίνει ότι η επεξηγηματική μεταβλητή που χρησιμοποιούμε είναι μάλλον ασθενής. Βλέπουμε ότι βρισκόμαστε μάλλον στη δεύτερη περίπτωση.

6. F-statistic. Οι τιμές της τελευταίας γραμμής αποτυπώνουν πιο αυστηρά την στατιστική εκτίμηση του μοντέλου. Η τιμή p-value μας λέει αν η εξίσωση που προκύπτει από το μοντέλο είναι πιο αποτελεσματική από μια τυχαία πρόβλεψη.

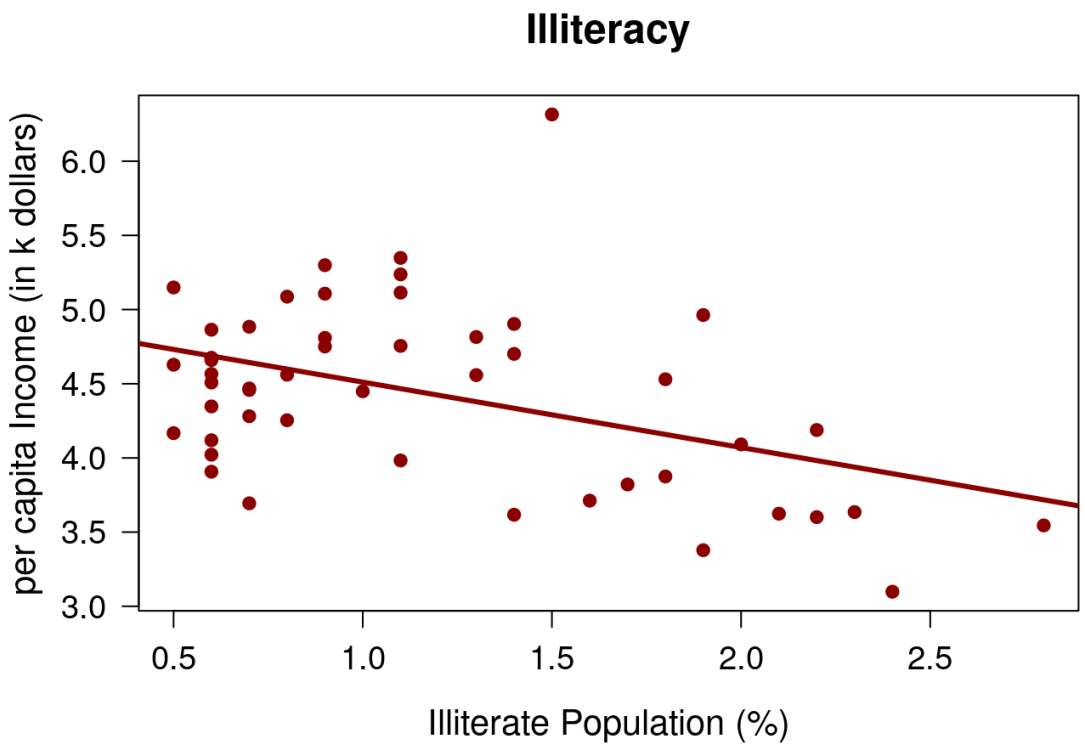


Γραφική αναπαράσταση μοντέλων παλινδρόμησης

Η συνάρτηση που μας επιτρέπει να κάνουμε κάτι τέτοιο λέγεται *abline*() και χρησιμοποιείται όπως η *lines*() μετά από την δημιουργία μιας γραφικής με την *plot*().

Hide

```
plot(state.data$Illiteracy, state.data$Income/1000, type="p", pch=19, col="dark red", ylab="per capita
Income (in k dollars)", xlab="Illiterate Population (%)", cex.lab=1.3, cex.axis=1.2, las=1, main="Illit
eracy", cex.main=1.5)
abline(lmfit.illit, col="dark red", lwd=3)
```



Στο διάγραμμα βλέπουμε ότι η γραμμική εξίσωση περνά σχετικά κοντά από τα σημεία αλλά και ότι τα σημεία από μόνα τους δεν έχουν μια κατανομή που να μας επιτρέπει να περιμένουμε ότι ένα γραμμικό μοντέλο θα έχει μεγάλη προβλεπτική ισχύ.

Γραφική αναπαράσταση μοντέλων παλινδρόμησης

Με απολύτως ανάλογο τρόπο ας εκτιμήσουμε ένα γραμμικό μοντέλο με βάση το ποσοστό αποφοίτων Λυκείου και ας δούμε την προβλεπτική του δυνατότητα αριθμητικά και γραφικά.

Hide

```
lmfit.hsgrad<-lm(Income/1000 ~ `HS Grad`, data=state.data)
summary(lmfit.hsgrad)
```

```
##
## Call:
## lm(formula = Income/1000 ~ `HS Grad`, data = state.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08313 -0.27741 -0.03415  0.24146  1.23817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.931105   0.462739   4.173 0.000125 ***
## `HS Grad`   0.047162   0.008616   5.474 1.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4871 on 48 degrees of freedom
## Multiple R-squared:  0.3843, Adjusted R-squared:  0.3715
## F-statistic: 29.96 on 1 and 48 DF,  p-value: 1.579e-06
```

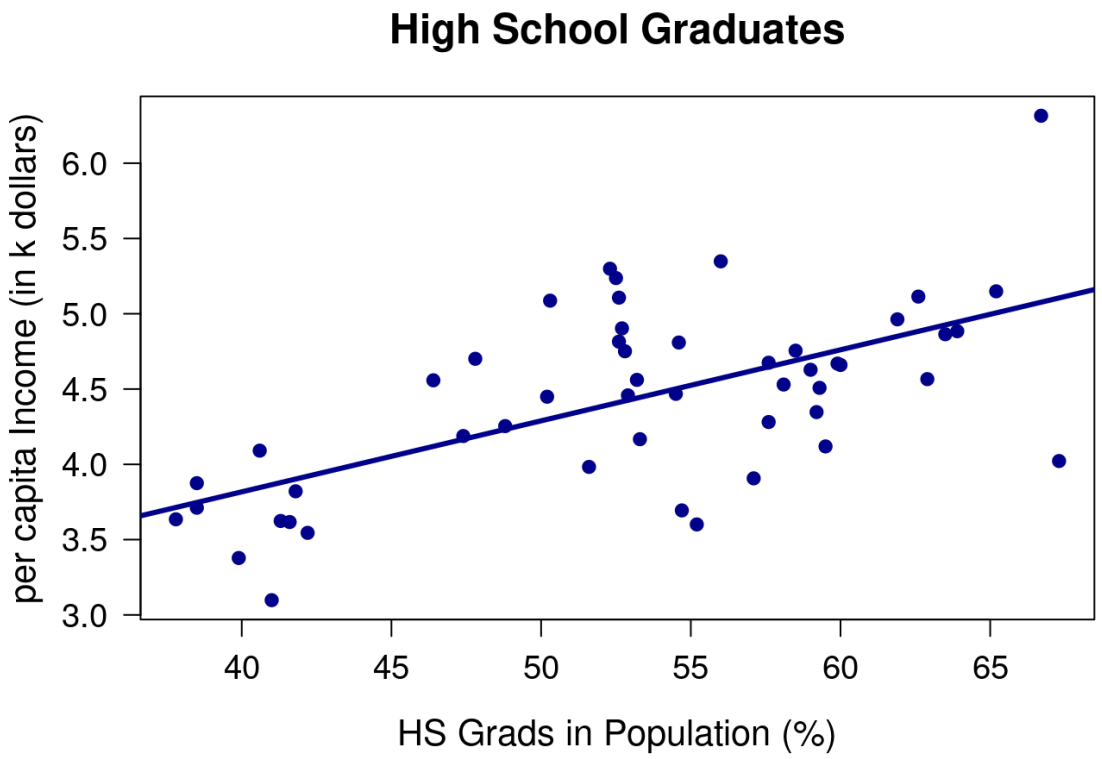
Βλέπουμε με μια ματιά ότι τόσο η τιμή p-value του coefficient HS Grad, όσο και η σχέση των R-squared τιμών αλλά και η τιμή p-value του F-statistic είναι ενδεικτικές καλύτερης προσαρμογής του μοντέλου στα δεδομένα, όπως φαίνεται εξάλλου και γραφικά.

Γραφική αναπαράσταση μοντέλων παλινδρόμησης

Βλέπουμε με μια ματιά ότι τόσο η τιμή p-value του coefficient HS Grad, όσο και η σχέση των R-squared τιμών αλλά και η τιμή p-value του F-statistic είναι ενδεικτικές καλύτερης προσαρμογής του μοντέλου στα δεδομένα, όπως φαίνεται εξάλλου και γραφικά.

Hide

```
plot(state.data$`HS Grad`, state.data$Income/1000, type="p", pch=19, col="dark blue", ylab="per capita
Income (in k dollars)", xlab="HS Grads in Population (%)", cex.lab=1.3, cex.axis=1.2, las=1, main="High
School Graduates", cex.main=1.5)
abline(lmfit.hsgrad, col="dark blue", lwd=3)
```



Από τη γραφική αναπαράσταση του μοντέλου βλέπουμε ότι το ποσοστό αποφοίτων είναι καλύτερος προβλεπτικός δείκτης για το ύψος του μέσου εισοδήματος από το ποσοστό αναλφαβητισμού.

Ύλη

Ανάλυση δεδομένων με την R

Κεφάλαιο 13