ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Ανάλυση Δεδομένων στη Λογιστική και Χρηματοοικονομική

Ενότητα 2η

Το Απλό Γραμμικό Υπόδειγμα Παλινδρόμησης

The classical (simple) linear regression model (CLRM)

Περιγραφή Ενότητας

- Απλή Γραμμική Παλινδρόμηση
- Εκτιμητές Ελαχίστων Τετραγώνων (OLS)
- Εκτίμηση Γραμμής Παλινδρόμησης
- Παράδειγμα ΥΑΚΣ
- Συντελεστής Προσδιορισμού και Συντελεστής Συσχέτισης
- Συσχέτιση και Αιτιότητα

Παλινδρόμηση (regression)

- Το μοντέλο της παλινδρόμησης είναι πιθανότατα το πιο σημαντικό εργαλείο στη φαρέτρα του ενός οικονομολόγου ερευνητή.

Όμως τι πραγματικά είναι η ανάλυση παλινδρόμησης;

- Η βασική ιδέα είναι η μοντελοποίηση (ποσοτικοποίηση) της σχέσης μεταξύ μίας δεδομένης μεταβλητής (της εξαρτημένης μεταβλητής – Y) και μίας ή περισσότερων μεταβλητών που αναμένεται να εξηγούν την εξαρτημένη μεταβλητή, οι οποίες ονομάζονται ανεξάρτητες μεταβλητές (X_1, X_2, κτλ).
- Το ποια μεταβλητή θέτουμε ως εξαρτημένη και το ποια(ες) ως ανεξάρτητες είναι μία πολύ σημαντική επιλογή η οποία θα πρέπει να υποστηρίζεται από κάποια οικονομική λογική (economic rationale). Όπως θα δούμε, οι βασικές υποθέσεις της απλής γραμμικής παλινδρόμησης περιπλέκουν περαιτέρω την απόφαση μας για το ποια μεταβλητή είναι εξαρτημένη και ανεξάρτητη(τες).

Σύμβολα

- Συνήθως η εξαρτημένη μεταβλητή ορίζεται με το σύμβολο y, ενώ οι ανεξάρτητες με τα σύμβολα x_1, x_2, ... , x_k όπου k ο αριθμός των ανεξάρτητων μεταβλητών.

- Τις μεταβλητές Y και X μπορεί να τις συναντήσουμε με όλα τα διαφορετικά ονόματα παρακάτω (ορολογία):

y	***x***
dependent variable	independent variable
regressand	regressors
effect variable	causal variable(s)
explained variable	explanatory variable(s)

Απλή Γραμμική Παλινδρόμηση (univariate regression)

- Ας υποθέσουμε ότι ένας ερευνητής έχει μία ιδέα (βασισμένη σε μία σκέψη που απορρέει από την οικονομική επιστήμη) για τη σχέση μεταξύ δύο μεταβλητών Y και X. Για παράδειγμα, έστω ότι η θεωρία προτείνει ότι μία αύξηση στη μεταβλητή X θα οδηγήσει σε μία αύξηση στη μεταβλητή Y. Σε αυτή την περίπτωση η Y εξαρτάται μόνο από τη μεταβλητή X.

- Για παράδειγμα:
  - Το πως οι αποδόσεις των περιουσιακών στοιχείων (π.χ. μετοχών) διακυμαίνονται σε σχέση με το συστηματικό τους ρίσκο (κίνδυνος αγοράς – market risk).
  - Η σχέση μεταξύ τιμών μετοχών και μερισμάτων.

Το μοντέλο της απλής γραμμικής παλινδρόμησης

Η σχέση μεταξύ εξαρτημένης (y_t) και ανεξάρτητης (x_t) θεωρείται ότι είναι γραμμική εάν επιλεγεί το μοντέλο της απλής γραμμικής παλινδρόμησης:

Population Y-Intercept (σταθερός όρος)

Population Slope (συντελεστής)

Random Error (διαταρακτικός όρος)

$$y_t = \alpha + \beta x_t + u_t$$

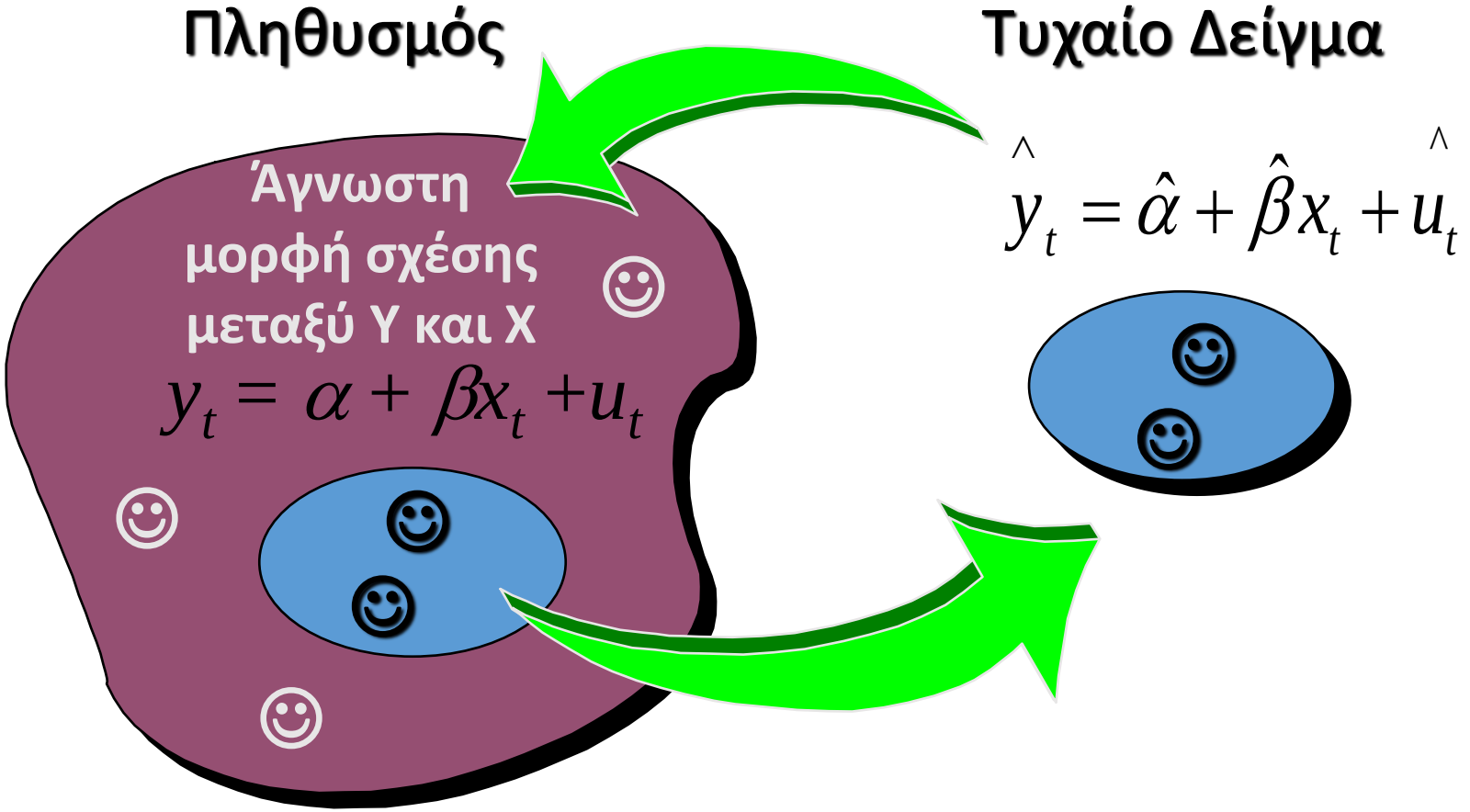
Dependent Variable (εξαρτημένη μεταβλητή)

Independent (Explanatory) Variable

(ανεξάρτητη ή επεξηγηματική μεταβλητή)

Πληθυσμός και δείγμα στην ανάλυση παλινδρόμησης

- Όπως είδαμε, ο πληθυσμός (population) είναι το σύνολο όλων των εξεταζόμενων ατόμων, αντικειμένων ή μετρήσεων σε μια έρευνα, ενώ δείγμα (sample) είναι κάθε υποσύνολο του πληθυσμού. Ένα τυχαίο δείγμα είναι ένα δείγμα που προέρχεται με ίση πιθανότητα επιλογής από ένα πληθυσμό.



Βρίσκοντας την «καλύτερη» γραμμή (εξίσωση) που περιγράφει τα δεδομένα

- Μπορούμε να διατυπώσουμε την εξίσωση μίας ευθείας γραμμής στη γενική της μορφή ως:

$$y=a+bx$$

αλλά σκοπός μας είναι να βρούμε την «καλύτερη» γραμμή που περιγράφει τα δεδομένα μας.

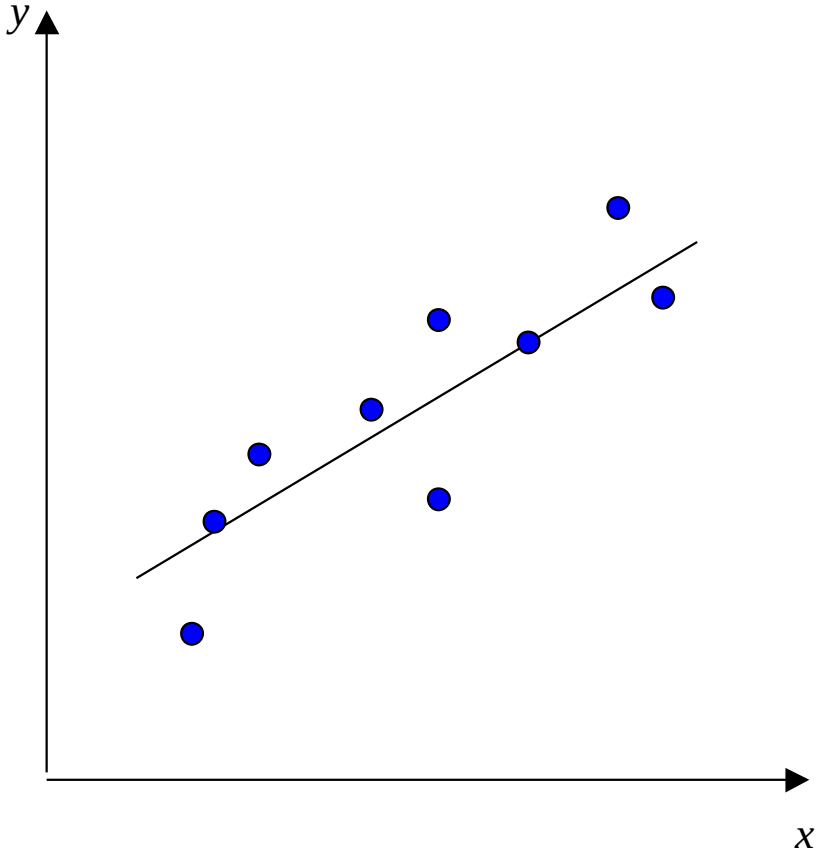
- Ωστόσο, η παραπάνω εξίσωση ($y=a+bx$) είναι τελείως ντετερμινιστική (deterministic), δηλαδή δεν αφήνει περιθώρια για τα δεδομένα μας να αποκλίνουν από τη γραμμική εξίσωση, κάτι το οποίο δεν είναι ρεαλιστικό.
- Για αυτό το λόγο προσθέτουμε ένα τυχαίο όρο u, στην παραπάνω εξίσωση, ως ακολούθως:

$$y_t = \alpha + \beta x_t + u_t$$

όπου $t = 1, 2, 3, 4, 5$

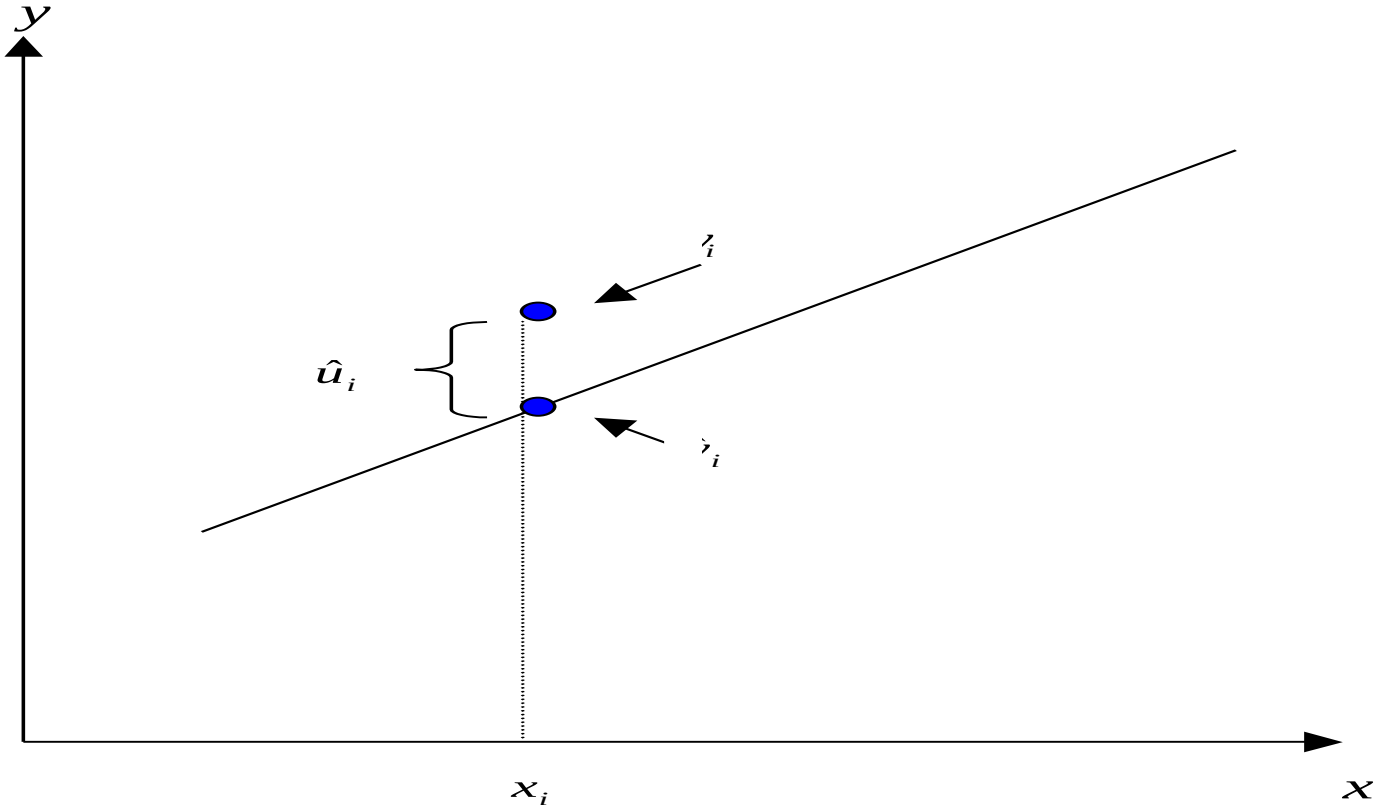
Προσδιορίζοντας τους εκτιμημένους συντελεστές (coefficients) της γραμμικής παλινδρόμησης

- Πως μπορούμε να προσδιορίσουμε τους εκτιμημένους συντελεστές (coefficients) α και β;
- Θα διαλέξουμε (εκτιμήσουμε) τους α και β κατά τέτοιο τρόπο ώστε οι κάθετες αποστάσεις κάθε σημείου των δεδομένων μας (μπλε κουκίδες) από την γραμμική εξίσωση (μαύρη γραμμή) να ελαχιστοποιούνται, αυτός είναι ο εκτιμητής ελαχίστων τετραγώνων – ordinary least squares OLS). Με αυτόν τον τρόπο η γραμμή περιγράφει τα δεδομένα μας όσο καλύτερα γίνεται:



Ordinary Least Squares (OLS) (1)

- Σύμφωνα με τον εκτιμητή OLS, αυτό που κάνουμε στην πράξη είναι να μετρήσουμε την απόσταση κάθε κουκίδας από τη γραμμή και να την υψώσουμε στο τετράγωνο. Στη συνέχεια ψάχνουμε εκείνη τη γραμμή η οποία ελαχιστοποιεί το άθροισμα όλων των (υψωμένων στο τετράγωνο) αποστάσεων των κουκίδων από τη γραμμή.
- Υψώνουμε τις αποστάσεις κάθε κουκίδας από τη γραμμή στο τετράγωνο γιατί σε διαφορετική περίπτωση οι αποστάσεις θα εξουδετέρωναν η μία την άλλη (θετικές και αρνητικές αποστάσεις κάθε κουκίδας από τη γραμμή).



y_t denotes the actual data point t

$\hat{y}_t$ denotes the fitted value from the regression line

$\hat{u}_t$ denotes the residual, equal to: $y_t - \hat{y}_t$

Ordinary Least Squares (OLS) (2)

- Ελαχιστοποιούμε το $\hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \hat{u}_4^2 + \hat{u}_5^2$ ή ισοδύναμα το $\sum_{t=1}^{5} \hat{u}_t^2$ άθροισμα των τετραγώνων των καταλοίπων (residual sum of squares).

- Με άλλα λόγια ελαχιστοποιούμε το $\sum (y_t - \hat{y}_t)^2$ που είναι ίσο με το $\sum \hat{u}_t^2$
- Οι εκτιμημένοι συντελεστές της εξίσωσης $\hat{\alpha}$ και $\hat{\beta}$ ορίζονται ως ακολούθως:

$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Παράμετροι από τον πληθυσμό και από το δείγμα

- Οι πραγματικοί συντελεστές α και β (όχι οι εκτιμημένοι) θεωρούνται αυτοί που πραγματικά γέννησαν τα δεδομένα που βλέπουμε και είναι άγνωστοι σε εμάς.

$$y_t = \alpha + \beta x_t + u_t$$

- Μπορούμε μόνο να τους προσεγγίσουμε εκτιμώντας τους συντελεστές χρησιμοποιώντας τα δεδομένα μας:

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$$

- Θέλουμε επίσης να γνωρίζουμε πόσο «καλοί» είναι οι εκτιμημένοι συντελεστές.

- Προσέξτε ότι είναι διαφορετική η έννοια του εκτιμητή, ο οποίος αναφέρεται στον τρόπο υπολογισμού ενός συντελεστή – δηλαδή τον τύπο που χρησιμοποιούμε, από την έννοια της εκτίμησης – δηλαδή την αριθμητική τιμή που παίρνει ο συντελεστής α και β.

Οι Υποθέσεις του Κλασσικού Γραμμικού Μοντέλου Παλινδρόμησης
Classical Linear Regression Model (CLRM)

- Το κλασσικό γραμμικό μοντέλο παλινδρόμησης στηρίζεται πάνω σε αρκετές υποθέσεις οι οποίες πρέπει να εκπληρώνονται από τις εκτιμήσεις μας, ώστε το μοντέλο να θεωρείται αξιόπιστο και να μπορεί να χρησιμοποιηθεί, π.χ. για την πρόβλεψη της εξαρτημένης μεταβλητής. Ιδιαίτερη έμφαση δίνεται στο διαταρακτικό όρο, καθώς τα δεδομένα για τις μεταβλητές Y και X είναι δεδομένα (δηλαδή είναι παρατηρήσεις). Ο διαταρακτικός όρος θα πρέπει να πληροί τις ακόλουθες υποθέσεις:

- Μαθηματικός Τύπος Ερμηνεία

1. $E(u_t) = 0$	Ο διαταρακτικός όρος έχει μέσο όρο μηδέν, *The errors have zero mean*
2. $Var(u_t) = \sigma^2$	Η διακύμανση του διαταρακτικού όρου είναι σταθερή και συγκεκριμένη για όλες τις τιμές της ανεξάρτητης μεταβλητής X, *The variance of the errors is constant and finite over all values of* x_t
3. $Cov(u_i,u_j)=0$	Οι διαταρακτικοί όροι είναι στατιστικά ανεξάρτητα μεταξύ τους (έχουν συνδιακύμανση μηδέν, *The errors are statistically independent of one another*
4. $Cov(u_t,x_t)=0$	Η συνδιακύμανση μεταξύ διαταρακτικών όρων και ανεξάρτητης μεταβλητής είναι μηδέν, *No relationship between the error and corresponding x variate*
5. u_t is normally distributed	Ο διαταρακτικός όρος κατανέμεται σύμφωνα με την κανονική κατανομή.

Παράδειγμα: Το Υπόδειγμα Αποτίμησης Περιουσιακών Στοιχείων (ΥΑΚΣ)
The Capital Asset Pricing Model (CAPM)

Το CAPM είναι ένα σημαντικό μοντέλο αποτίμησης περιουσιακών στοιχείων για την επιστήμη των χρηματοοικονομικών. Σύμφωνα με το CAPM οι επενδυτές πρέπει να ανταμείβονται (έξτρα απόδοση) μόνο για το συστηματικό κίνδυνο που αναλαμβάνουν (δηλαδή τον κίνδυνο αγοράς). Η «ευαισθησία» κάθε περιουσιακού στοιχείου στον κίνδυνο αγοράς μπορεί να μετρηθεί με το συντελεστή βήτα (beta). Αν μία μετοχή έχει συντελεστή βήτα πάνω από την μονάδα θεωρείται επιθετική και αν είναι κάτω από την μονάδα ως αμυντική, ενώ το ίδιο το χαρτοφυλάκιο της αγοράς (ο γενικός δείκτης) έχει βήτα ίσο με τη μονάδα. Ως χαρτοφυλάκιο της αγοράς συνήθως χρησιμοποιούμε τις αποδόσεις του γενικού δείκτη ενός χρηματιστηρίου, μία υπόθεση που ελέγχεται.

$$r_{i,t} - r_{f,t} = a_i + \beta_i \left(r_{m,t} - r_{f,t} \right) + u_{i,t}$$

$$r_{i,t} - r_{f,t} \; is \; the \; risk \; premium \; of \; stock \; i \; at \; time \; t$$

$$r_{m,t} - r_{f,t} \; is \; the \; risk \; premium \; of \; the \; market \; portfolio$$

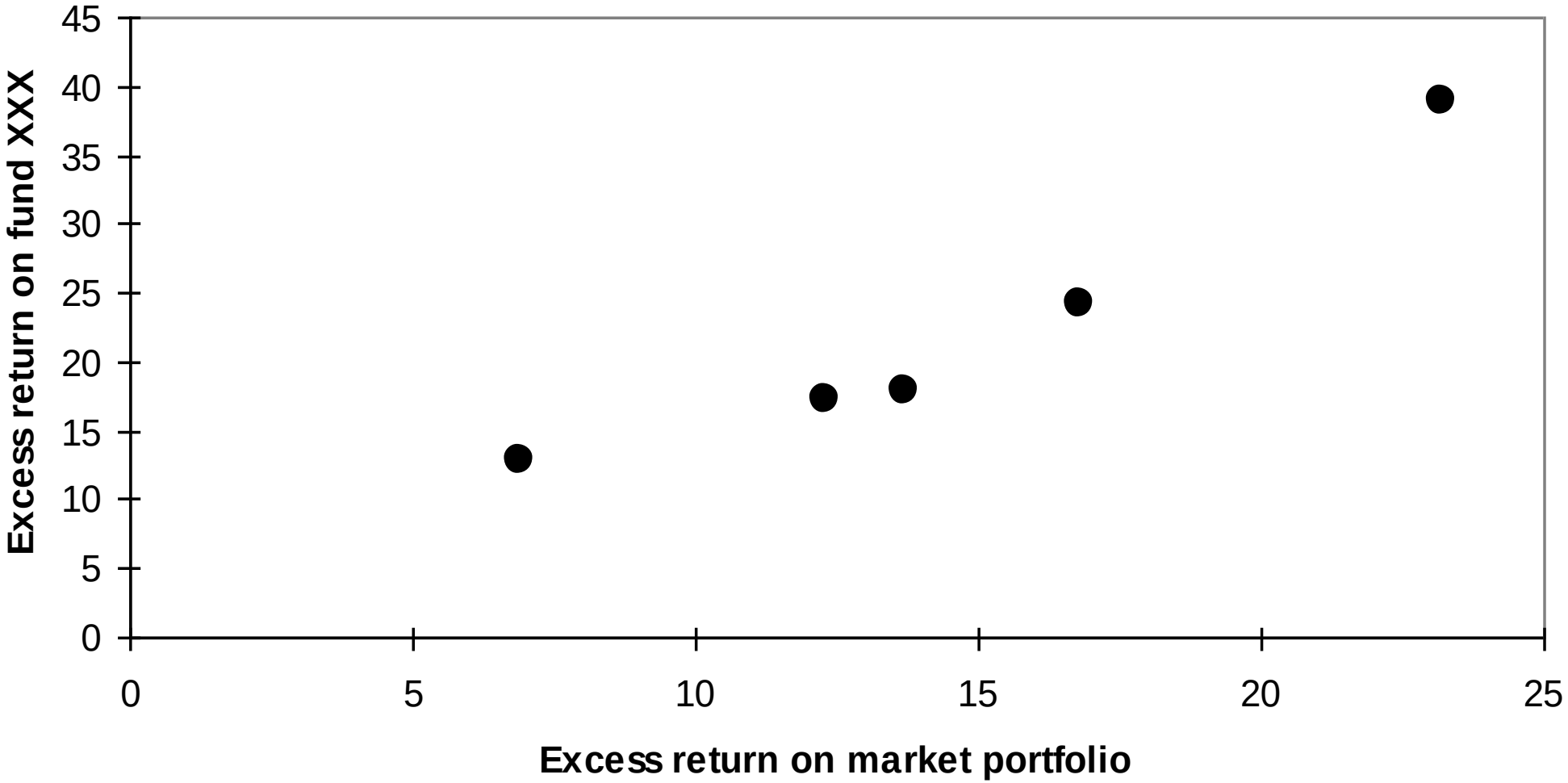
Το Υπόδειγμα Αποτίμησης Περιουσιακών Στοιχείων (ΥΑΚΣ)

- Ας υποθέσουμε ότι έχουμε τα ακόλουθα δεδομένα για τις υπεραποδόσεις (excess returns) ενός χαρτοφυλακίου ενός αμοιβαίου κεφαλαίου (mutual fund) και τις υπεραποδόσεις για τον γενικό δείκτη (market index)

Year, t	Excess return $= r_{XXX,t} - rf_t$	Excess return on market index $= rm_t - rf_t$
1	17.8	13.7
2	39.0	23.2
3	12.8	6.9
4	24.2	16.8
5	17.2	12.3

- Κοιτώντας τα δεδομένα έχουμε κάποια διαίσθηση ότι ο συντελεστής βήτα θα είναι θετικός. Ένα πρώτο βήμα είναι να δημιουργήσουμε ένα γράφημα μεταξύ των δύο αποδόσεων (του αμοιβαίου κεφαλαίου και της αγοράς), δηλαδή ένα scatter plot.

Scatter Plot



Εκτίμηση του ΥΑΚΣ με ένα απλό γραμμικό μοντέλο παλινδρόμησης

- Χρησιμοποιώντας τις πέντε παρατηρήσεις που έχουμε για τις δύο αποδόσεις (του αμοιβαίου κεφαλαίου και του γενικού δείκτη) μπορούμε να εκτιμήσουμε τους συντελεστές α και β μίας απλής γραμμικής παλινδρόμησης ως:

$$\hat{y}_t = -1.74 + 1.64 x_t$$

- Μία πρόβλεψη που κάνει το παραπάνω οικονομετρικό μοντέλο είναι για παράδειγμα αν κάποιος αναλυτής σας πει ότι περιμένει την απόδοση του χαρτοφυλακίου της αγοράς να είναι μεγαλύτερη από το επιτόκιο χωρίς ρίσκο (risk free rate) κατά 20%, τότε ποια περιμένετε να είναι η απόδοση ενός αμοιβαίου κεφαλαίου ;
- Μπορούμε να απαντήσουμε εάν θέσουμε το x_t ίσο με 20% και λύσουμε ως προς y_t την παραπάνω γραμμική εξίσωση:

$$\hat{y}_i = -1.74 + 1.64 \times 20 = 31.06$$

Συντελεστής Προσδιορισμού (R^2)

- Ο πιο εύκολος τρόπος εκτίμησης την ερμηνευτικής δύναμης ενός γραμμικού μοντέλου είναι ο ***συντελεστής προσδιορισμού (coefficient of determination)*** που συνήθως συμβολίζεται με R^2. Ο συντελεστής αυτός μετρά πόση διακύμανση της εξαρτημένης μεταβλητής κατάφεραν να ερμηνεύσουν οι ανεξάρτητες μεταβλητές. Ουσιαστικά είναι το πιο απλό μέτρο που μετρά την ικανότητα ενός συνόλου παραγόντων να ερμηνεύσουν ένα φαινόμενο. Ο συντελεστής προσδιορισμού R^2 είναι ο λόγος της διακύμανσης των εκτιμημένων τιμών της εξαρτημένης μεταβλητής προς τη διακύμανση των πραγματικών τιμών της εξαρτημένης μεταβλητής και υπολογίζεται ως εξής:

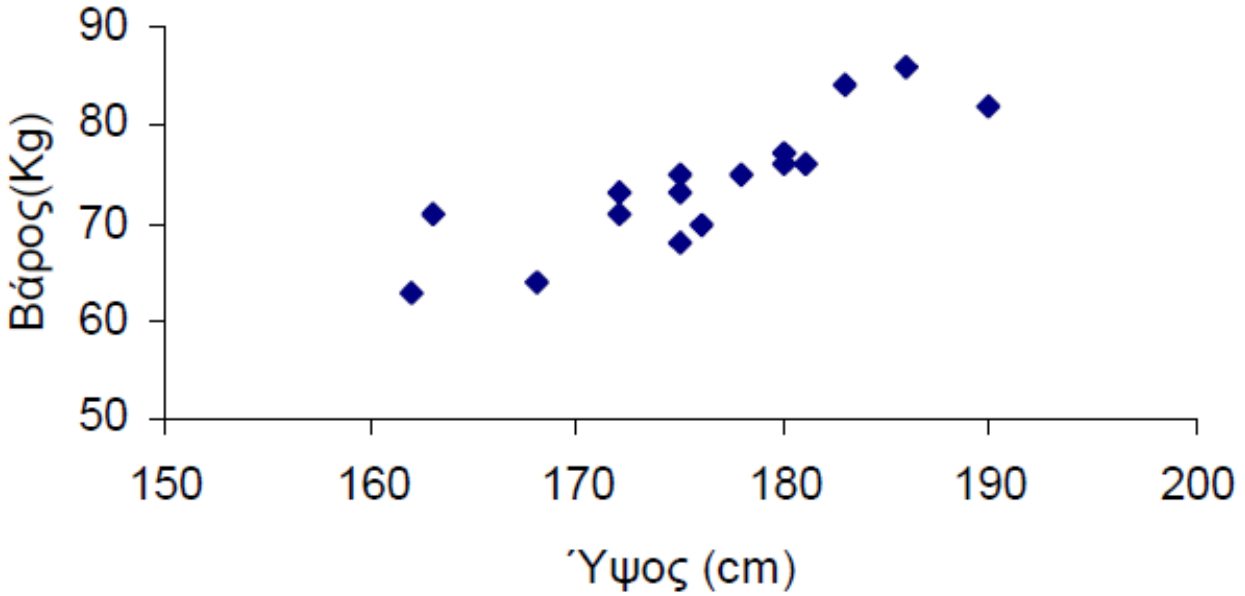
$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- όπου n είναι ο αριθμός των παρατηρήσεων, y_i είναι οι πραγματικές τιμές της εξαρτημένης μεταβλητής Y, $\bar{y}$ είναι η μέση τιμής της μεταβλητής Y και $\hat{y}_i$ είναι οι εκτιμημένες τιμές της Y.
- Οι τιμές του συντελεστή προσδιορισμού R^2 κυμαίνονται από το 0 ως το 1 και όσο η τιμή πλησιάζει προς το 1 τόσο καλύτερη προσαρμογή έχει το μοντέλο. Για παράδειγμα $R^2 = 1$ σημαίνει ότι οι ερμηνευτικές μεταβλητές εξηγούν το 100% της διακύμανσης της εξαρτημένης μεταβλητής και άρα έχουμε ένα τέλειο μοντέλο Στην πράξη ο συντελεστής προσδιορισμού R^2 θεωρείται ικανοποιητικός ή όχι ανάλογα με την εμπειρική εφαρμογή.

Τι είναι η συσχέτιση; (Correlation)

- Η συσχέτιση είναι ένα μέτρο της ισχύος της σχέσης μεταξύ δύο μεταβλητών. Ο συντελεστής συσχέτισης ποσοτικοποιεί τον βαθμό αλλαγής σε μια μεταβλητή με βάση την αλλαγή στην άλλη μεταβλητή. Στα στατιστικά στοιχεία, η συσχέτιση συνδέεται με την έννοια της εξάρτησης, η οποία είναι η στατιστική σχέση μεταξύ δύο μεταβλητών.
- Ο συντελεστής συσχέτισης του Pearsons ή απλά ο συντελεστής συσχέτισης r είναι μια τιμή μεταξύ -1 και 1 ($-1 \leq r \leq +1$). Είναι ο συντελεστής συσχέτισης που χρησιμοποιείται πιο συχνά και ισχύει μόνο για μια γραμμική σχέση μεταξύ των μεταβλητών.
  - Εάν $r = 0$, δεν υπάρχει σχέση και εάν $r \geq 0$, η σχέση είναι άμεσα ανάλογη. δηλ. η τιμή μιας μεταβλητής αυξάνεται με την αύξηση της άλλης.
  - Εάν $r \leq 0$, η σχέση είναι αντιστρόφως ανάλογη. δηλ. μία μεταβλητή μειώνεται καθώς η άλλη αυξάνεται.
- Εξαιτίας της γραμμικότητας του, ο συντελεστής συσχέτισης r μπορεί επίσης να χρησιμοποιηθεί για να διαπιστωθεί η παρουσία γραμμικής σχέσης μεταξύ των μεταβλητών.

- *Στον πίνακα που ακολουθεί φαίνονται οι παρατηρήσεις που πήραμε για το ύψος και το βάρος 16 εργατών μιας βιομηχανίας.*
- Ύψος (cm) 183 162 172 181 180 168 176 180 190 175 178 175 186 172 175 163
- Βάρος (Kg) 84 63 71 76 77 64 70 76 82 68 75 73 86 73 75 71
- Από το *διάγραμμα διασποράς* φαίνεται ότι οι εργάτες στο δείγμα που έχουν μεγαλύτερο ύψος έχουν και μεγαλύτερο βάρος. Φαίνεται, δηλαδή, να υπάρχει μια ανάλογη σχέση μεταξύ του ύψους και του βάρους των εργατών. Πόσο ισχυρή είναι όμως αυτή η συσχέτιση;
- Ο *δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson* συμβολίζεται με *r*, ενώ ο *πληθυσμιακός συντελεστής γραμμικής συσχέτισης του Pearson* ορίζεται ανάλογα και συμβολίζεται με ρ.

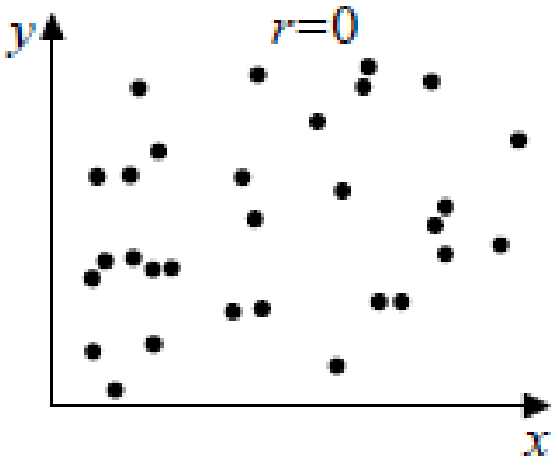


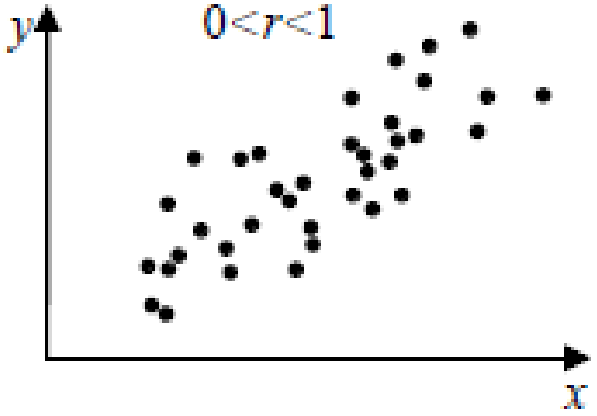
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

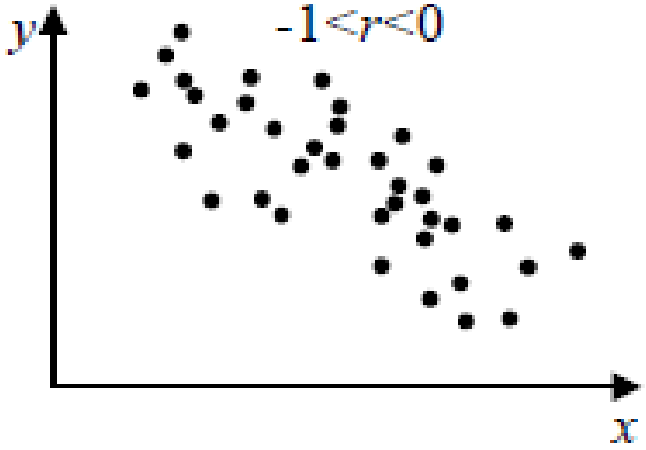
όπου,

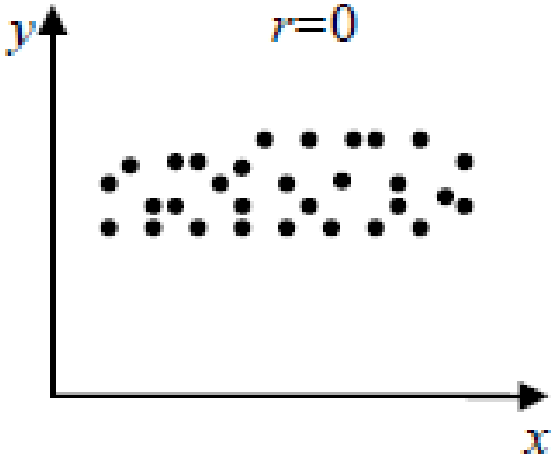
$$s_{xy} = Cov(X,Y) = \frac{\sum_{i=1}^{\nu}(\chi_i - \overline{\chi})\cdot(y_i - \overline{y})}{\nu - 1} = \frac{\sum_{i=1}^{\nu}\chi_i y_i - \nu \cdot \overline{\chi} \cdot \overline{y}}{\nu - 1}$$

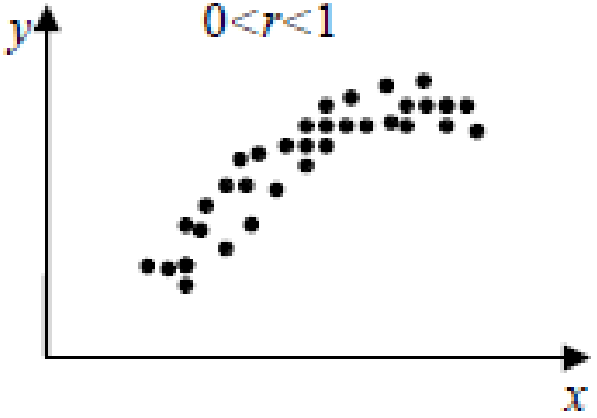
$$s_x = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{\nu}(\chi_i - \overline{\chi})^2} \quad \text{και} \quad s_y = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{\nu}(y_i - \overline{y})^2}$$

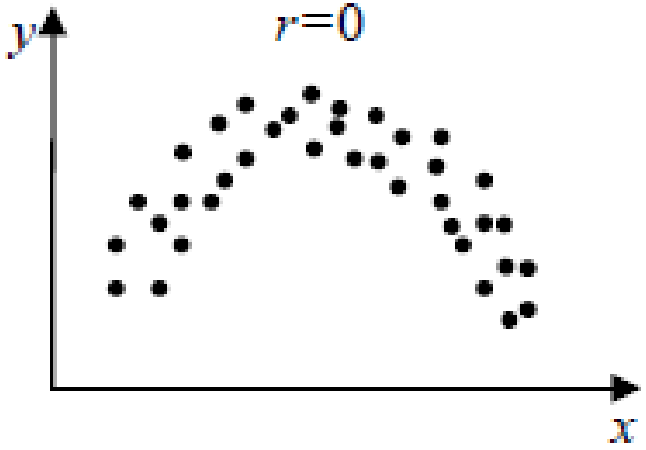

y
r=0
x


y
0<r<1
x


y
-1<r<0
x


y
r=0
x


y
0<r<1
x


y
r=0
x

- Ποια είναι η βασική ιδέα στον ορισμό του συντελεστή του Pearson; Στο παράδειγμα, το μέσο ύψος είναι 176 cm και το μέσο βάρος 74Kg. Παρατηρείστε ότι οι εργάτες που έχουν ύψος πάνω από το μέσο ύψος έχουν (στις περισσότερες περιπτώσεις) και βάρος πάνω από το μέσο βάρος. Ανάλογα, οι εργάτες που έχουν ύψος κάτω από το μέσο ύψος έχουν (στις περισσότερες περιπτώσεις) και βάρος κάτω από το μέσο βάρος.

Παράδειγμα-1

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) . (y_i - \bar{y})$
1	4	-2	4	4	16	-8
2	2	-1	2	1	4	-2
3	0	0	0	0	0	0
4	-2	1	-2	1	4	-2
5	-4	2	-4	4	16	-8
$\sum x_i = 15$	$\sum y_i = 0$			$\sum (x_i - \bar{x})^2 = 10$	$\sum (y_i - \bar{y})^2 = 40$	$\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = -20$

$$\bar{x} = 3,\ \bar{y} = 0$$

$$r = \frac{\sum_{i=1}^{5} (\chi_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{5} (\chi_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{5} (y_i - \bar{y})^2}} = \frac{-20}{\sqrt{10} \cdot \sqrt{40}} = -1$$

Παράδειγμα –2

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	-2	1	4	-2
3	0	9	0	0
5	1	25	1	5
7	3	49	9	21
9	5	81	25	45
10	6	100	36	60
12	8	144	64	96
13	10	169	100	130
$\sum x_i = 60$	$\sum y_i = 31$	$\sum x_i^2 = 578$	$\sum y_i^2 = 239$	$\sum x_i \cdot y_i = 355$

$\bar{x} = 7,5 \;\; και \;\; \bar{y} = 3,9$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu \cdot \bar{\chi} \cdot \bar{y}}{\sqrt{\sum_{i=1}^{\nu} \chi_i^2 - \nu \cdot \bar{\chi}^2} \cdot \sqrt{\sum_{i=1}^{\nu} y_i^2 - \nu \cdot \bar{y}^2}} = \frac{355 - 8 \cdot 7,5 \cdot 3,9}{\sqrt{578 - 8 \cdot 7,5^2} \cdot \sqrt{239 - 8 \cdot 3,9^2}} = 0,99$$

Συσχέτιση vs. Παλινδρόμηση

- Η παλινδρόμηση δίνει τη μορφή της σχέσης μεταξύ δύο τυχαίων μεταβλητών και η συσχέτιση δίνει τον βαθμό ισχύος της σχέσης.
- Η ανάλυση παλινδρόμησης παράγει μια συνάρτηση παλινδρόμησης, η οποία βοηθά στην παρέκταση και την πρόβλεψη των αποτελεσμάτων, ενώ η συσχέτιση μπορεί να παρέχει πληροφορίες μόνο για ποια κατεύθυνση μπορεί να αλλάξει.
- Τα ακριβέστερα μοντέλα γραμμικής παλινδρόμησης δίδονται από την ανάλυση, εάν ο συντελεστής συσχέτισης είναι υψηλότερος. ($| r | \geq 0.8$)
- Για παράδειγμα υψηλή θετική συσχέτιση μεταξύ δύο μεταβλητών σημαίνει ότι οι δύο μεταβλητές κινούνται προς την ίδια κατεύθυνση και με παρόμοια ένταση. Ωστόσο, στην παλινδρόμηση η ανεξάρτητη μεταβλητή (X) εξηγεί τις διακυμάνσεις της εξαρτημένης μεταβλητής (Y) κατά ένα στατιστικά σημαντικό τρόπο.

Προσοχή: άλλο συσχέτιση, άλλο αιτιότητα

- Η συσχέτιση -στην πιο γνωστή της εκδοχή, χάρις στον Pearson-είναι ένα μέτρο του κατά πόσο δύο ποσότητες μπορούν να παρατηρηθούν να έχουν γραμμική εξάρτηση μεταξύ τους. Είναι μια πολύ συνηθισμένη ποσότητα για την έκθεση των αποτελεσμάτων επιστημονικών ερευνών, ειδικά, αλλά όχι αποκλειστικά, στις κοινωνικές επιστήμες. Οι ερευνητές προσπαθούν να αποδείξουν την ύπαρξη συσχέτισης μεταξύ δύο φαινομένων ως πρωταρχικό βήμα για την εξέταση μιας πιθανής σχέσης αιτιότητας μεταξύ τους.
- Δεν υπάρχει φυσικά τίποτα το λάθος στη μέτρηση της συσχέτισης. Το πρόβλημα έγκειται στην ερμηνεία των αποτελεσμάτων. Εάν για παράδειγμα εκτιμήσω μία υψηλή θετική συσχέτιση μεταξύ της κατανάλωσης σοκολάτας ανά χώρα και του αριθμού των υποψηφίων της χώρας για βραβείο Νόμπελ, θα έπρεπε να συμπεράνω ότι η κατανάλωση σοκολάτας σε κάνει εξυπνότερο; Ή πως το να κερδίζεις βραβεία Νόμπελ σε κάνει να τρως περισσότερες σοκολάτες…;

Διάρθρωση Δεδομένων (structure of dataset)

- Σε κάθε γραμμική παλινδρόμηση οι μεταβλητές Y και X μπορούν να έχουν τα ακόλουθα χαρακτηριστικά:
  - Παλινδρόμηση χρονολογικών σειρών. Σε αυτή την περίπτωση οι δύο μεταβλητές (ή περισσότερες) αλλάζουν μόνο σε σχέση με το χρόνο. Για παράδειγμα τα μέσα έξοδα για φαγητό ανά τυπικό νοικοκυριό στην Ελλάδα (Y) και το μέσο εισόδημα ενός τυπικού νοικοκυριού στην Ελλάδα (X).
  - Και οι δύο αυτές μεταβλητές μπορούν μόνο να αλλάζουν σε σχέση με το χρόνο, και η συχνότητα των δεδομένων μπορεί να ποικίλει, π.χ. εβδομαδιαία, μηνιαία, τριμηνιαία, εξαμηνιαία, ετήσια, κ.ο.κ.

Διάρθρωση Δεδομένων (structure of dataset)

- Όμως για σκεφτείτε για παράδειγμα ότι θέλετε να μελετήσετε ποιες μεταβλητές εξηγούν τα κόστη πωληθέντων (Selling and Administrative Expenses, SGA) ; Πιθανές μεταβλητές μπορεί να είναι:
  - οι πωλήσεις, το σύνολο του ενεργητικού και άλλες μεταβλητές οι οποίες αλλάζουν για κάθε μία εταιρεία που μελετάτε στο δείγμα σας και ταυτόχρονα για κάθε στιγμή στο χρόνο (π.χ. τριμηνιαία ή ετήσια δεδομένα), firm-level variables.
  - το ΑΕΠ της χώρας, ο πληθωρισμός κτλ. Αυτές οι μεταβλητές είναι οι ίδιες για όλες τις εταιρείες μίας χώρας και αλλάζουν μόνο με τον χρόνο (π.χ. τριμηνιαία ή ετήσια δεδομένα), macro-level variables.
  - και πως μπορούμε να λάβουμε υπόψη ότι οι διαφορετικοί κλάδοι της οικονομίας μίας χώρας επίσης μπορεί να παίζουν ρόλο ; (μέσο περιθώριο κέρδους για κάθε κλάδο). Industry-level variable.

Πως να διαμορφώσουμε τα δεδομένα μας

- Για τους παραπάνω λόγους, θα πρέπει σε κάθε ερευνητικό ερώτημα να επιλέξουμε τις μεταβλητές Y και (X_1, X_2, κτλ.) πολύ προσεκτικά, δηλαδή βασισμένοι σε ένα οικονομικό σκεπτικό (economic rationale) αλλά φυσικά και στη διαθεσιμότητα των δεδομένων που υπάρχουν (ή που έχουμε πρόσβαση σε αυτά).
- Στη συνέχεια θα πρέπει να αναρωτηθούμε αν η μεταβλητή Y είναι μία firm-level, industry-level, macro-level variable και στη συνέχεια να καθορίσουμε τις X_1, X_2, κτλ. αναλόγως.
- Έτσι λοιπόν η γραμμική ανάλυση παλινδρόμησης κυρίως αναφέρεται σε χρονολογικές σειρές (όπου όλες οι μεταβλητές είναι macro-level) ή σε panel data (δηλαδή ένα συνδυασμό χρονολογικών σειρών με διαστρωματικά στοιχεία).
- Ερώτηση: Αν κάποιος εξετάσει την ανεργία (Y) σε σχέση με το ΑΕΠ (X), αλλά έχει δεδομένα όχι για μία χώρα αλλά για όλες τις χώρες της ΕΕ, τότε αυτό είναι μία παλινδρόμηση χρονολογικών σειρών (macro-level) ;
- Απάντηση: Όχι απαραίτητα. Το ιδανικό θα ήταν να εξετάσει όλες τις χώρες ταυτόχρονα, δημιουργώντας ένα panel dataset. Θα μπορούσε όμως να τρέξει μία παλινδρόμηση για κάθε μία χώρα ξεχωριστά (λιγότερο σημαντικά αποτελέσματα).

APPENDIX

Deriving the OLS Estimator

- But $\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$, so let $L = \sum_t (y_t - \hat{y}_t)^2 = \sum_i (y_t - \hat{\alpha} - \hat{\beta} x_t)^2$

- Want to minimise L with respect to (w.r.t.) $\hat{\alpha}$ and $\hat{\beta}$, so differentiate L w.r.t. $\hat{\alpha}$ and $\hat{\beta}$

$$\frac{\partial L}{\partial \hat{\alpha}} = -2\sum_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \qquad (1)$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2\sum_t x_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \qquad (2)$$

- From (1), $\sum_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \Leftrightarrow \sum y_t - T\hat{\alpha} - \hat{\beta} \sum x_t = 0$

- But $\sum y_t = T\bar{y}$ and $\sum x_t = T\bar{x}$.

Deriving the OLS Estimator (cont'd)

- So we can write $T\bar{y} - T\hat{\alpha} - T\hat{\beta}\bar{x} = 0$ or $\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$ (3)
- From (2), $\sum_t x_t (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0$ (4)
- From (3), $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ (5)
- Substitute into (4) for $\hat{\alpha}$ from (5),

$$\sum_t x_t (y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta} x_t) = 0$$

$$\sum_t x_t y_t - \bar{y}\sum x_t + \hat{\beta}\bar{x}\sum x_t - \hat{\beta}\sum x_t^{\,2} = 0$$

$$\sum_t x_t y_t - T\bar{y}\bar{x} + \hat{\beta}T\bar{x}^2 - \hat{\beta}\sum x_t^{\,2} = 0$$

Deriving the OLS Estimator (cont'd)

- Rearranging for $\hat{\beta}$,

$$\hat{\beta}(T\overline{x}^2 - \sum x_t^2) = T\overline{y}\overline{x} - \sum x_t y_t$$

- So overall we have

$$\hat{\beta} = \frac{\sum x_t y_t - T\overline{x}\overline{y}}{\sum x_t^2 - T\overline{x}^2} \text{and} \hat{\alpha} = \bar{y} - \hat{\beta}\overline{x}$$

- This method of finding the optimum is known as ordinary least squares.