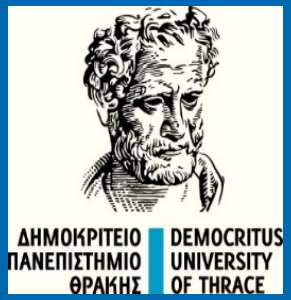


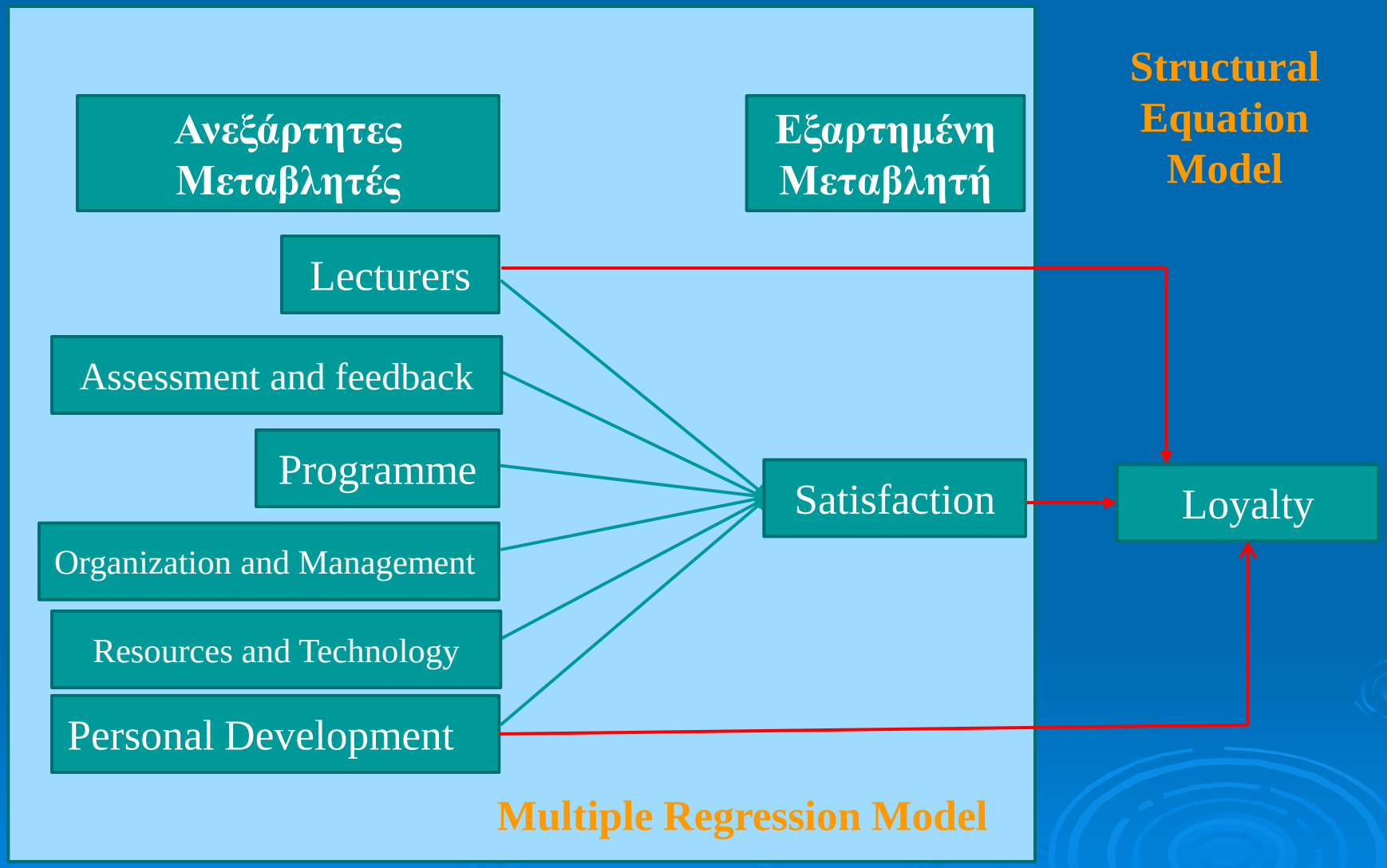

# Δημοκρίτειο Πανεπιστήμιο Θράκης

# Μεταπτυχιακό Πρόγραμμα: Διοίκηση Τουριστικών Επιχειρήσεων και Οργανισμών για Στελέχη (Executive MBA in Tourism)

Dr. Efstathios Dimitriadis
Mathematic
Ph.D in Applied Statistics
M.Sc in Statistics and Demography
M.Sc in Quality Assurance

Καβάλα, 2024

**Εννοιολογικό Πλαίσιο**: Παράγοντες που επηρεάζουν θετικά την ικανοποίηση των φοιτητών από τις σπουδές τους



# Ανάλυση Παλινδρόμησης

Η ***ανάλυση παλινδρόμησης*** στοχεύει στην οικοδόμηση σχέσεων μεταξύ μιας μοναδικής ***εξαρτημένης μεταβλητής*** ή ***μεταβλητής απόκρισης*** και μιας ή περισσότερων ***ανεξάρτητων ή προγνωστικών μεταβλητών*** και είναι μια από τις ευρύτερα χρησιμοποιούμενες μεθόδους στην ανάλυση δεδομένων.

Οι Τρεις βασικές χρήσεις της ανάλυσης παλινδρόμησης είναι:

1. Ο καθορισμός της ισχύος των ανεξάρτητων μεταβλητών
2. Η πρόβλεψη ενός αποτελέσματος και
3. Η πρόβλεψη τάσεων.

# Υποθέσεις της Παλινδρόμησης

Η γραμμική παλινδρόμηση βασίζεται στο παρακάτω σετ υποθέσεων:

1. ***Πλήθος Δεδομένων***. Η ιδανική αναλογία είναι 20 περιπτώσεις για κάθε ανεξάρτητη μεταβλητή. Η μικρότερη αποδεκτή αναλογία πρέπει να είναι 5:1
2. ***Γραμμικότητα***. Γραμμικότητα (linearity) σημαίνει ότι υπάρχει μια ευθεία σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής. Αυτή η υπόθεση είναι σημαντική επειδή η ανάλυση παλινδρόμησης ελέγχει μόνο μια γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής. Οποιαδήποτε μη γραμμική σχέση μεταξύ της ανεξάρτητης μεταβλητής και της εξαρτημένης μεταβλητής αγνοείται.
3. ***Κανονικότητα- Normality***. Τα δεδομένα των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής κατανέμονται κανονικά.

4. ***Συγραμμικότητα ή πολυσυγραμμικότητα*** .

Η πολυσυγραμμικότητα (multicollinearity) είναι μια κατάσταση στην οποία οι ανεξάρτητες μεταβλητές συσχετίζονται πολύ ($|r| \geq 0.7$). Οι υψηλοί διμεταβλητοί συσχετισμοί είναι εύκολο να εντοπιστούν με απλή εκτέλεση συσχετίσεων μεταξύ των ανεξάρτητων μεταβλητών σας. Εάν έχετε υψηλούς συσχετισμούς διμεταβλητών, το πρόβλημά σας επιλύεται εύκολα διαγράφοντας μία από τις δύο μεταβλητές.

Η συγραμμικότητα και η πολυσυγραμμικότητα εξετάζονται με τους δείκτες Tolerance και Variance Inflation Factor (V.I.F). Ένας δείκτης (V.I.F=1/Tolerance) V.I.F <5 υποδεικνύει ότι δεν υπάρχει κανένα πρόβλημα της γραμμικότητας ή της πολυσυγραμμικότητας.

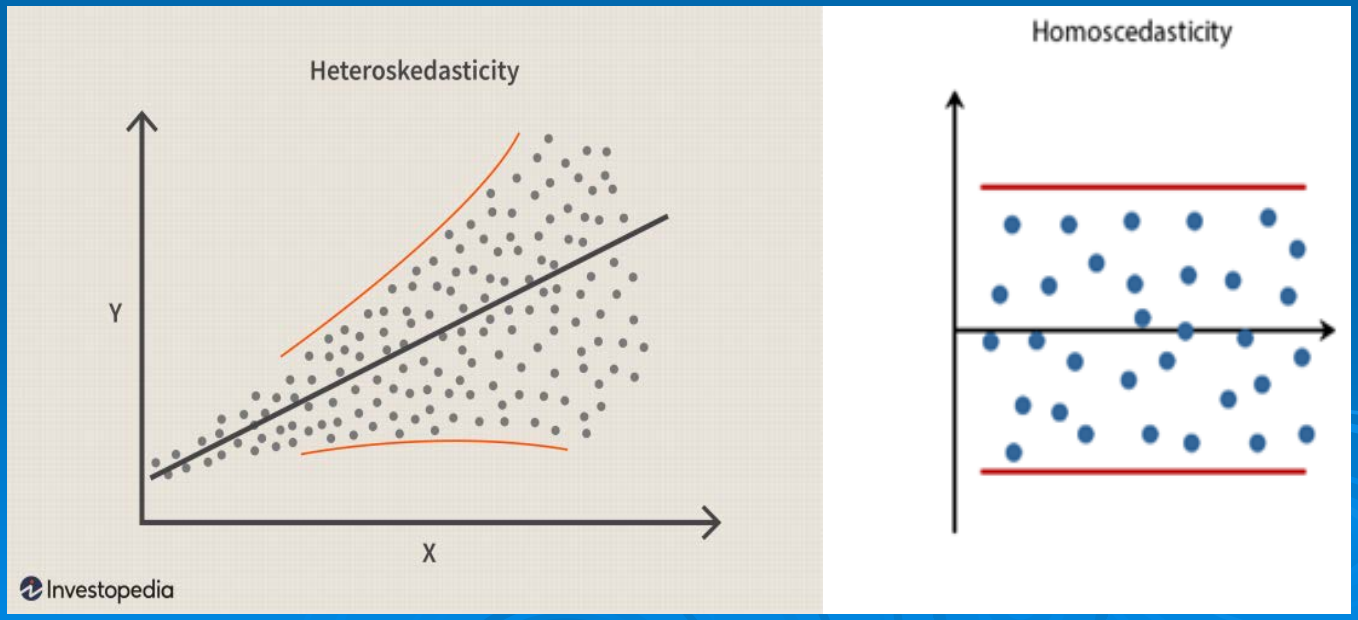
5. ***Αυτοσυσχέτιση.*** Η αυτοσυσχέτιση- Autocorrelation (που ονομάζεται επίσης σειριακή συσχέτιση) εμφανίζεται όταν συσχετίζονται οι παρατηρήσεις των σφαλμάτων σε μια παλινδρόμηση. Το κατάλληλο τεστ για την Αυτό-συσχέτιση είναι το τεστ Durbin-Watson. Η τιμή του κυμαίνεται μεταξύ 0 και 4. Μια τιμή D.W κοντά στο 2 δεν δείχνει αυτόματη συσχέτιση. Όταν η τιμή είναι κάτω από το 2, δείχνει θετική αυτοσυσχέτιση και τιμή μεγαλύτερη από 2 δείχνει αρνητική σειριακή συσχέτιση.

Ωστόσο, οι τιμές κάτω του 1 και άνω 3 προκαλούν ανησυχία και ενδέχεται να καταστήσουν την ανάλυσή σας άκυρη. Συχνά τα δεδομένα που χρησιμοποιούνται για μελέτες παλινδρόμησης στις επιχειρήσεις και τα οικονομικά συλλέγονται με την πάροδο του χρόνου. Δεν είναι ασυνήθιστο η τιμή του y στο χρόνο t, που υποδηλώνεται από το $y_t$, να σχετίζεται με την τιμή του y στην προηγούμενη χρονική περίοδο που υποδηλώνεται από το $y_{t-1}$.

Ο έλεγχος Durbin-Watson για αυτοσυσχέτιση χρησιμοποιεί τα κατάλοιπα: $d_i = y_i - \hat{y}_i$

Durbin- Watson Statistic test: $$D.W = \frac{\sum_{t=2}^{n} (d_t - d_{t-1})^2}{\sum_{t=1}^{n} d_t^2}$$

***6. Ομοσκεδαστικότητα.*** Η παραδοχή της ομοσκεδαστικότητας (homoskedasticity) είναι ότι τα ***κατάλοιπα*** είναι περίπου ίσα για όλες τις προβλεπόμενες τιμές. Ένας άλλος τρόπος σκέψης είναι ότι η μεταβλητότητα στις προβλεπόμενες τιμές για τις ανεξάρτητες μεταβλητές σας είναι η ίδια σε όλες τις τιμές της εξαρτημένης μεταβλητής. Οι τιμές δεδομένων για εξαρτημένες και ανεξάρτητες μεταβλητές έχουν ***ίσες διακυμάνσεις***.



# Ακραίες Τιμές- Extreme Values

Οι ***ακραίες τιμές*** στις στατιστικές αναλύσεις είναι απομακρυσμένες τιμές που δεν φαίνεται να ταιριάζουν με την πλειονότητα ενός συνόλου δεδομένων. Εάν δεν αφαιρεθούν, αυτές οι ακραίες τιμές μπορεί να έχουν μεγάλη επίδραση σε τυχόν συμπεράσματα που θα μπορούσαν να εξαχθούν από τα εν λόγω δεδομένα.

Ως ακραίες τιμές καθορίζονται όλες οι τιμές οι οποίες είναι:

Κάτω από Q1-1,5*IQR και

Πάνω από Q3+1,5*IQR

Q1 και Q3 είναι το πρώτο και το τρίτο τεταρτημόριο και IQR είναι το ενδοτεταρτομοριακό έυρος.

***!!! Στην ανάλυση παλινδρόμησης, μια τιμή χαρακτηρίζεται ως ακραία αν η τιμή των τυποποιημένων καταλοίπων είναι μικρότερη του -2 ή μεγαλύτερη του +2.***

# Επηρεάζουσες Παρατηρήσεις- Influential Observations

Μερικές φορές, στην ανάλυση παλινδρόμησης, μία ή περισσότερες παρατηρήσεις έχουν ισχυρή επιρροή στα αποτελέσματα. Αυτές οι παρατηρήσεις ονομάζονται ***Επηρεάζουσες Παρατηρήσεις***.

Η εκτιμώμενη γραμμή παλινδρόμησης από αρνητική κλίση μπορεί να αλλάξει σε θετική κλίση εάν η επηρεάζουσα παρατήρηση αποκλειστεί από τα δεδομένα. Είναι σαφές ότι αυτή η παρατήρηση έχει πολύ μεγαλύτερη επιρροή στον προσδιορισμό της γραμμής της εκτιμώμενης παλινδρόμησης από οποιαδήποτε άλλη.

Μια επηρεάζουσα παρατήρηση μπορεί να είναι μια ακραία τιμή (μια παρατήρηση με τιμή y που αποκλίνει ουσιαστικά από την τάση), μπορεί να αντιστοιχεί σε μια τιμή x μακριά από τη μέση τιμή ή μπορεί να προκαλείται από έναν συνδυασμό των δύο.

Οι παρατηρήσεις με ακραίες τιμές για την ανεξάρτητη μεταβλητή ονομάζονται ***σημεία υψηλής μόχλευσης***. Η μόχλευση μιας παρατήρησης καθορίζεται από το πόσο μακριά είναι οι τιμές των ανεξάρτητων μεταβλητών από τις μέσες τιμές τους. Η μόχλευση της παρατήρησης i συμβολίζεται με $h_i$ και μπορεί να υπολογιστεί χρησιμοποιώντας τον τύπο:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Στην περίπτωση της απλής γραμμικής παλινδρόμησης, αν η $h_i$, για μια παρατήρηση είναι μεγαλύτερη από τον λόγο 6/n ($h_i$>6/n), αυτή η παρατήρηση ονομάζεται ***σημείο υψηλής μόχλευσης.*** Στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης χρησιμοποιείται ο κανόνας $h_i$>3(p+1)/n για τον καθορισμό των ***επηρεαστικών παρατηρήσεων***.

## Προσοχή !!!!!

Πολλές φορές χρησιμοποιώντας την ***μόχλευση*** για να καθορίσουμε τις επηρεαστικές παρατηρήσεις, μια παρατήρηση μπορεί να χαρακτηριστεί ως υψηλής μόχλευσης ενώ δεν είναι απαραίτητα επιδραστική στα αποτελέσματα της παλινδρόμησης. Έτσι, σε κάποιες περιπτώσεις η χρήση μόνο της μόχλευσης για τον εντοπισμό επηρεαστικών παρατηρήσεων μπορεί να μας οδηγήσει σε λάθος συμπεράσματα.

Για τον λόγο το μέτρο ***Cook's distance*** είναι το καταλληλότερο, καθώς χρησιμοποιεί την μόχλευση και το κατάλοιπο της i παρατήρησης, για τον καθορισμό του αν μια παρατήρηση είναι ή όχι επηρεαστική.

$$D_i = \frac{(y_i - \hat{y}_1)^2}{(p-1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Αν $D_i>1$ τότε η i παρατήρηση είναι επηρεαστική.

# Τύποι Γραμμικής Παλινδρόμησης

**1. Απλή Γραμμική Παλινδρόμηση**

(μια μόνο ανεξάρτητη μεταβλητή)

**2. Πολλαπλή Γραμμική Παλινδρόμηση**

(δύο ή περισσότερες ανεξάρτητες μεταβλητές)

# 1.Απλή Γραμμική Παλινδρόμηση

(μια μόνο ανεξάρτητη μεταβλητή)

# Απλή Γραμμική Παλινδρόμηση

- Η απλή γραμμική παλινδρόμηση περιλαμβάνει μια ανεξάρτητη μεταβλητή και μία εξαρτημένη μεταβλητή.
- Η σχέση μεταξύ των δύο μεταβλητών προσεγγίζεται από μια ευθεία γραμμή.

## Μοντέλο Απλής Γραμμικής Παλινδρόμησης

- Το μοντέλο της απλής γραμμικής παλινδρόμησης είναι:

$$y = \beta_0 + \beta_1 \chi + \varepsilon$$

όπου:

$\beta_0$ και $\beta_1$ ονομάζονται ***παράμετροι*** του μοντέλου,

$\varepsilon$ είναι μια τυχαία μεταβλητή ονομαζόμενη ***όρος σφάλματος***

Υποθέτουμε ότι ο όρος σφάλματος $\varepsilon$ ***ακολουθεί την*** $N(0, \sigma^2)$

# Συνάρτηση Απλής Γραμμικής Παλινδρόμησης

- Η συνάρτηση της απλής γραμμικής παλινδρόμησης δίνεται από την σχέση:

$$E(y) = \beta_0 + \beta_1 \chi$$

  - Η γραφική απεικόνιση της συνάρτησης είναι μια ευθεία γραμμή
  - $\beta_0$ είναι η τομή της ευθείας με τον άξονα y.
  - $\beta_1$ είναι η κλίση της ευθείας.
  - $E(y)$ είναι η αναμενόμενη τιμή της y για δεδομένη τιμή της χ.

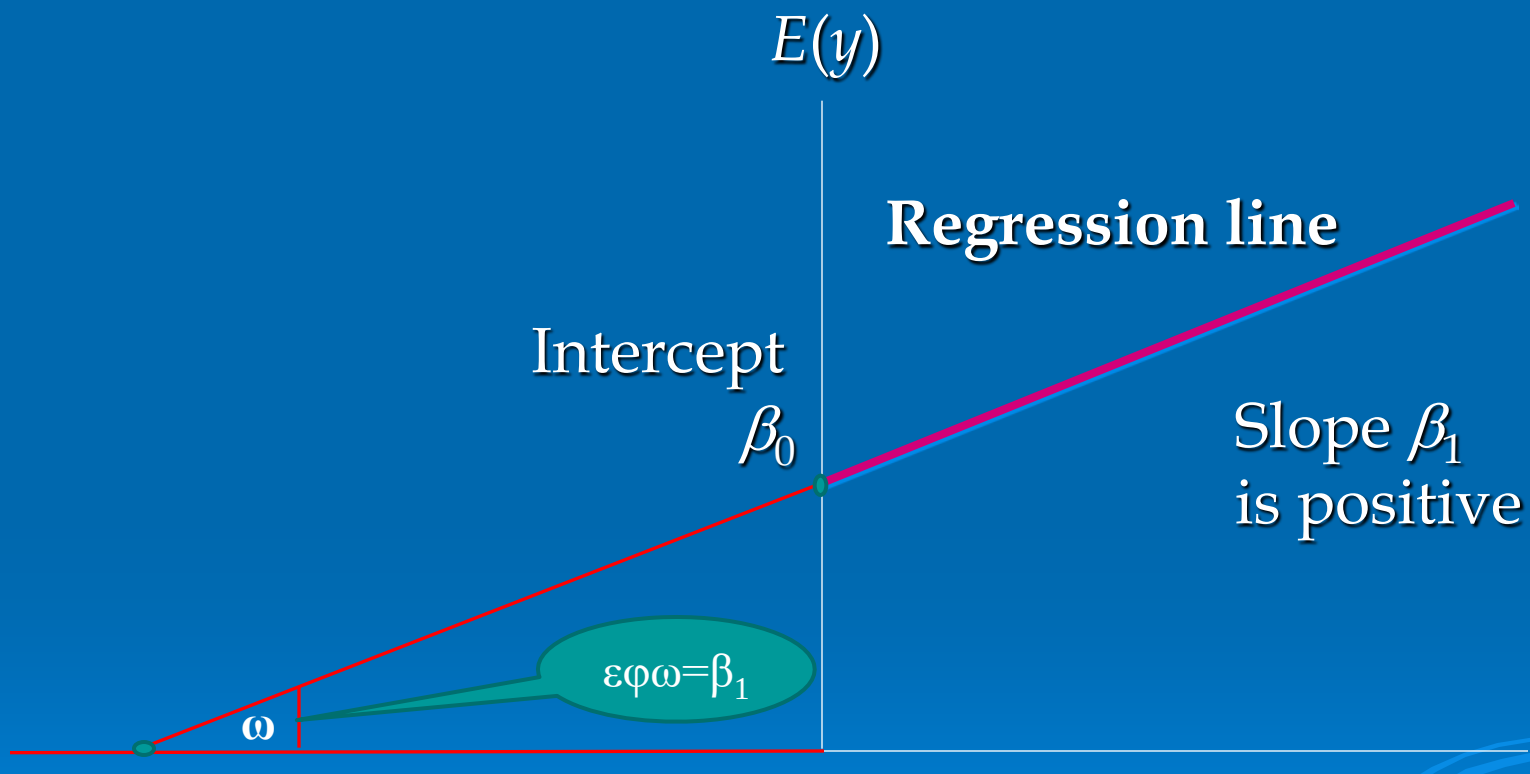
# Εκτιμώμενη συνάρτηση απλής γραμμικής παλινδρόμησης

Εκτιμώμενη συνάρτηση απλής γραμμικής παλινδρόμησης

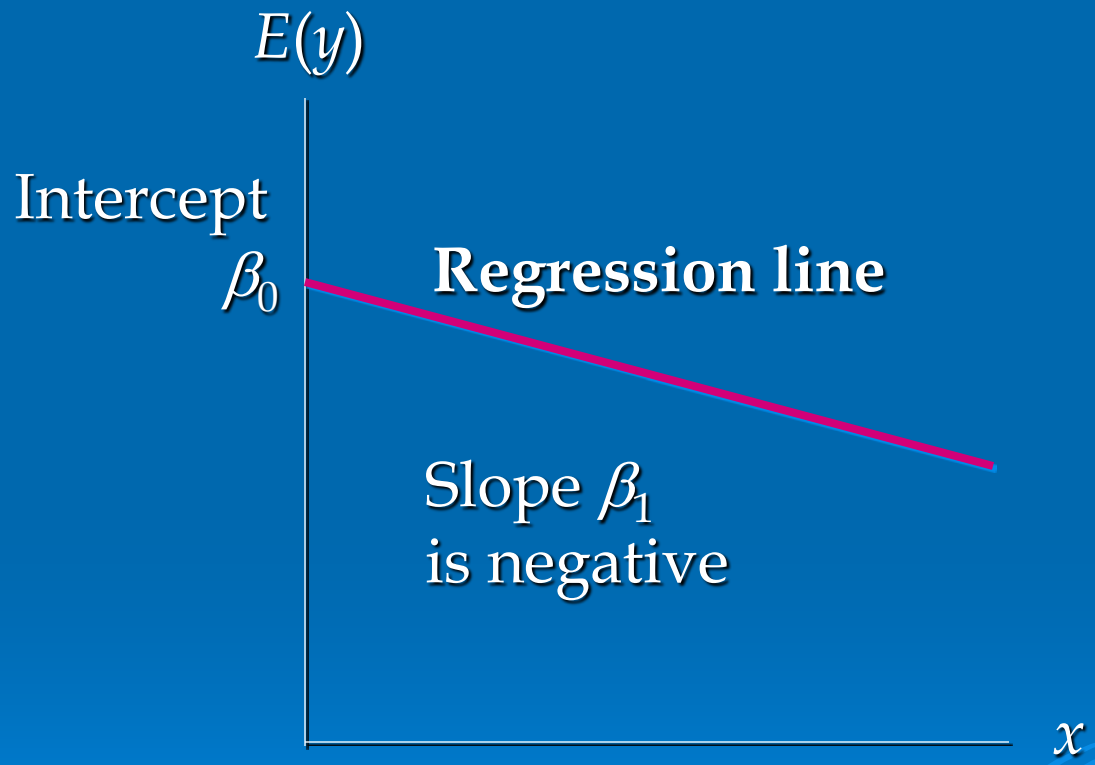
$$\widehat{Y}_i = \widehat{b}_0 + \widehat{b}_1 X_i$$

- Το γράφημα ονομάζεται εκτιμώμενη γραμμή παλινδρόμησης
- $\widehat{b}_0$ είναι η τομή της ευθείας με τον άξονα y.
- $\widehat{b}_1$ είναι η κλίση της ευθείας.
- $\hat{y}$ είναι η Εκτιμώμενη της μεταβλητής y για δεδομένη τιμή της μεταβλητής χ.

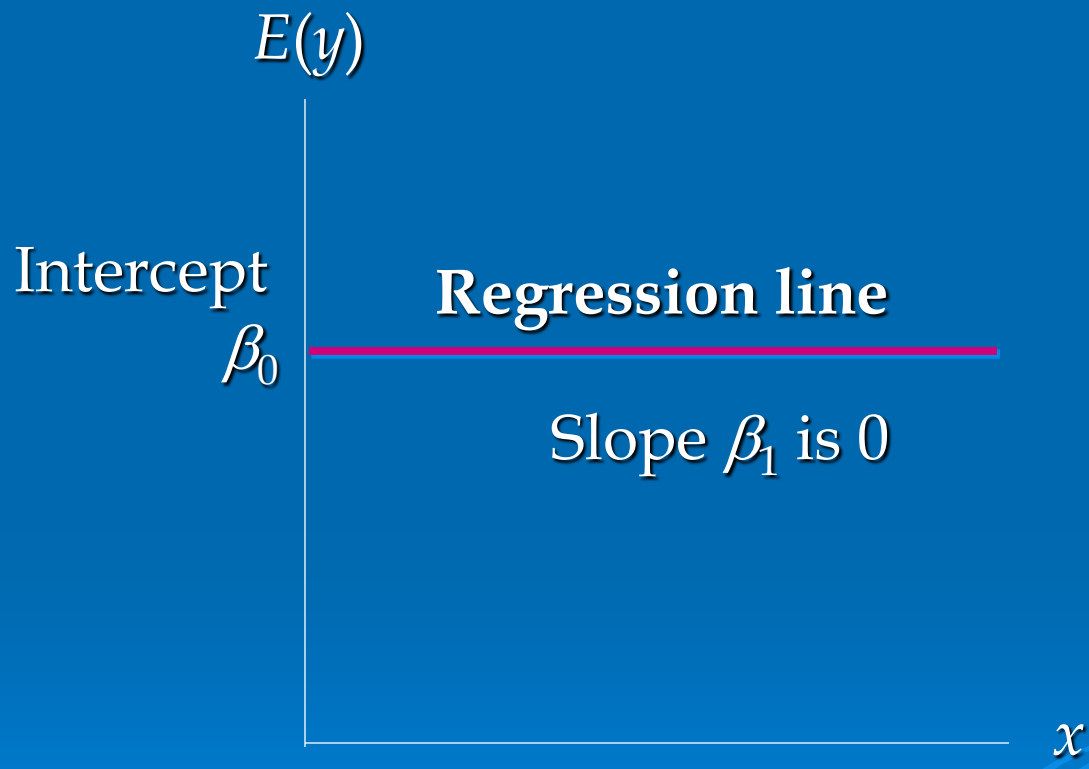
- **Positive Linear Relationship**



- **Negative Linear Relationship**



- **No Relationship**



# Μοντέλο Ελαχίστων Τετραγώνων

$$\widehat{Y}_i = \widehat{b}_0 + \widehat{b}_1 X_i$$

***Λάθος ή Κατάλοιπο***: $d_i = y_i - \hat{y}_i$

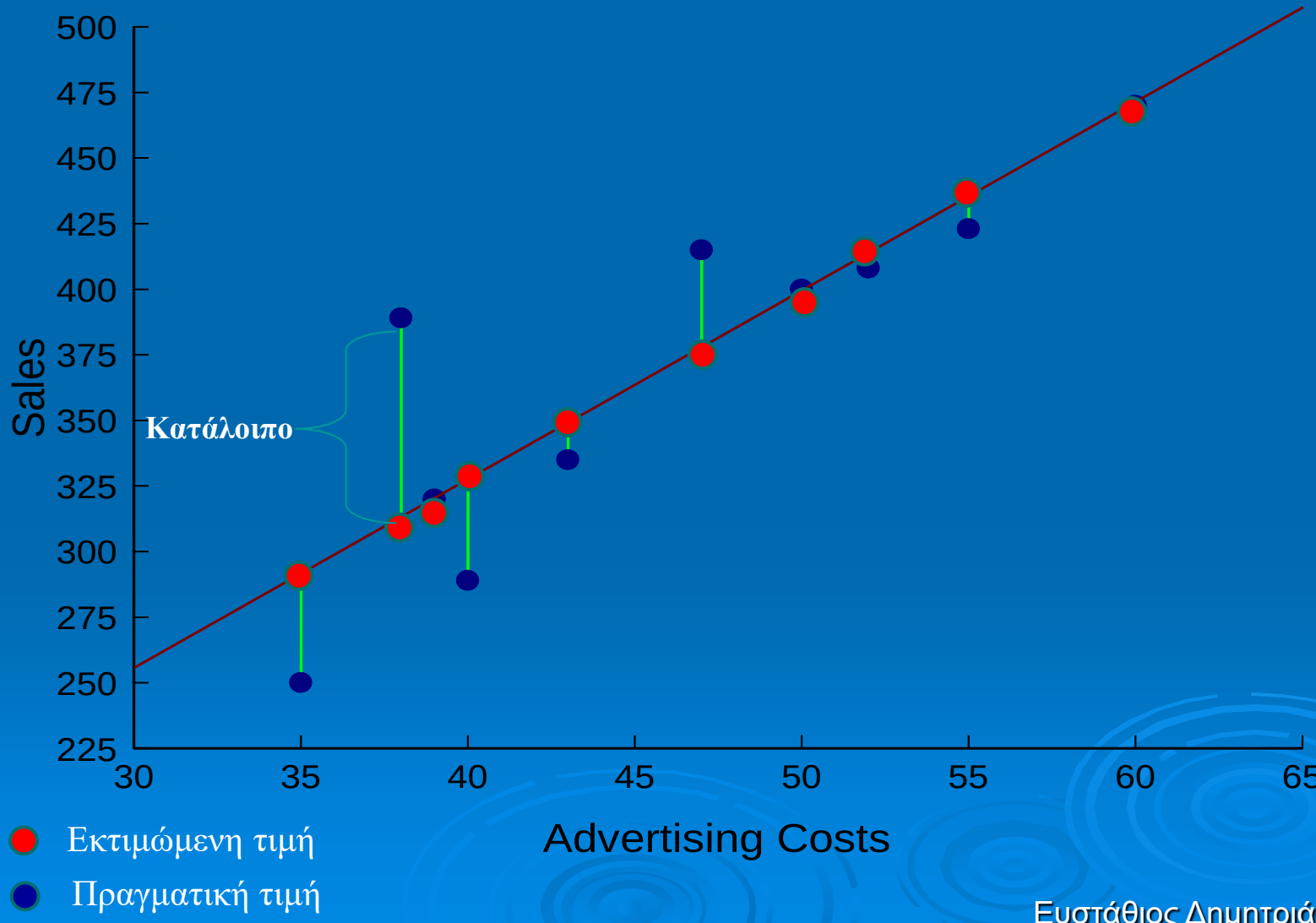
Η τυχαία μεταβλητή d, ονομάζεται κατάλοιπο και είναι η διαφορά μεταξύ της πραγματικής τιμής και της εκτιμώμενης τιμής της μεταβλητής y για δεδομένη τιμή της μεταβλητής χ.

***Κριτήριο Ελαχίστων τετραγώνων:*** $\sum d_i^2 = \min$

***Η κλίση της ευθείας δίνεται από τον τύπο:*** $\widehat{b}_1 = \dfrac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2}$

***Το σημείο τομής με τον άξονα y από τον τύπο:*** $\widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x}$

# Γραμμή Παλινδρόμησης και Κατάλοιπα



**Μέσο τετραγωνικό σφάλμα:** $\sigma^2 = \frac{\sum d_i^2}{n} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n}$

**Τυπική απόκλιση:** $\sigma = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n}}$

**ή** $\sigma = \sqrt{\frac{\sum Y_i^2 - \hat{b}_0 \sum Y_i - \hat{b}_1 \sum X_i \cdot Y_i}{n}}$

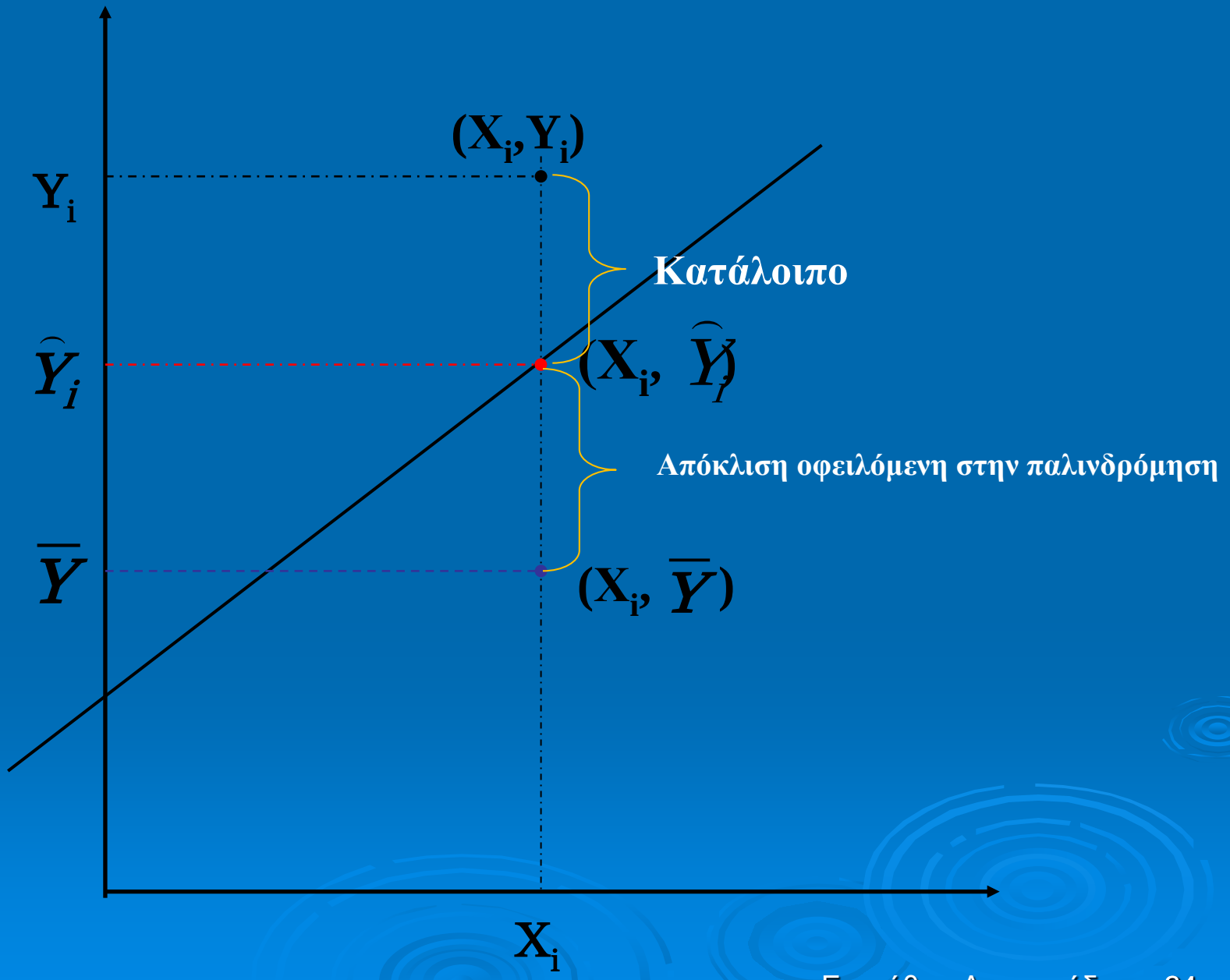
# Συντελεστής Προσδιορισμού
# R- Square

**Χρησιμοποιείται για να εκφράσει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από τη διακύμανση των ανεξάρτητων μεταβλητών.**

**Παίρνει τιμές από 0 έως 1**

- **Η τιμή 1 δείχνει τέλεια ερμηνεία.**
- **Η τιμή 0 δείχνει ότι δεν υπάρχει σχέση**

***!!!! Πρακτικά είναι αδύνατον να πάρει την τιμή 1 καθώς η συνάρτηση είναι στοχαστική***


$(X_i, Y_i)$
$Y_i$
Κατάλοιπο
$\widehat{Y}_i$
$(X_i, \widehat{Y}_i)$
Απόκλιση οφειλόμενη στην παλινδρόμηση
$\overline{Y}$
$(X_i, \overline{Y})$
$X_i$

# Συντελεστής Προσδιορισμού

**RSS=** $\sum(\hat{Y}_i - \bar{Y})^2$ **sum of square due to regression**

**ESS=** $\sum(Y_i - \hat{Y}_i)^2$ **sum of square due to error**

**TSS =** $\sum(Y_i - \bar{Y})^2$ **total sum of square**

**TSS=RSS+ESS**

$$R^2 = \frac{RSS}{TSS} = \frac{TSS - ESS}{TSS} = 1 - \frac{ESS}{TSS}$$

# Συντελεστής Προσδιορισμού

$$R^2 = \frac{\sigma^2_{\hat{Y}}}{\sigma^2_Y} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

$$\sigma^2_Y = \frac{\sum(Y_i - \bar{Y})^2}{n} \qquad \sigma^2_{\hat{Y}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{n}$$

$$R^2 = \frac{\sigma^2_{\hat{Y}}}{\sigma^2_Y} = \frac{\sigma^2_Y - \sigma^2}{\sigma^2_Y} \Leftrightarrow R^2 = 1 - \frac{\sigma^2}{\sigma^2_Y}$$

Adjusted R- square: $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$

Where k is the number of independent variables