7-3-2023

$Y = b_0 + b_1 X + \varepsilon$ , $\varepsilon \sim N(0,\sigma^2)$

$H_0: b_1 = 0$ $H_1: b_1 \neq 0$. έλεγχος στατιστικής σημαντικότητας της $b_1$ (όλου του μοντέλου)

t-test

n-2 β.ε.

Εκτίμηση/Πρόβλεψη για νέες τιμές του X.

Για $X = x_0$ $Y = ?$

$Y = b_0 + b_1 x_0 + \varepsilon$ $\varepsilon \sim N(0,\sigma^2)$

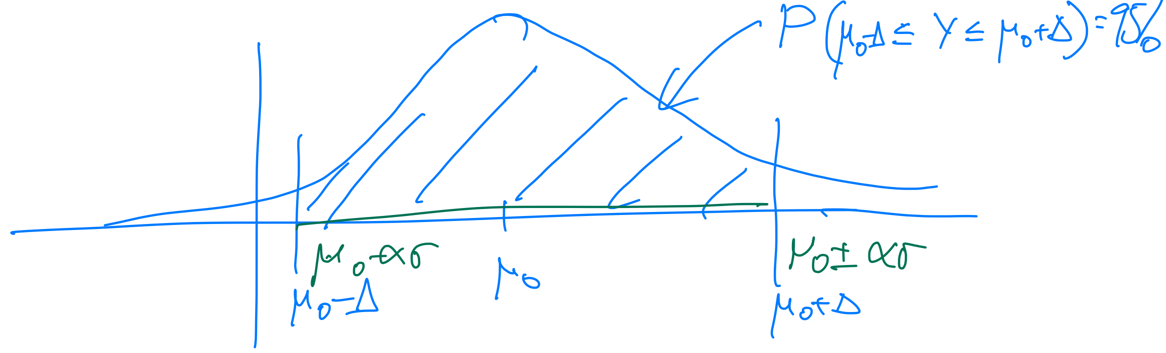
(1) $\Leftrightarrow E(Y) = b_0 + b_1 x_0$ $\Rightarrow$ $\boxed{E(Y) = ? \text{ για } X = x_0 ?}$ Εκτίμηση

(2) Αν έχω ένα μεμονωμένο άτομο με $X = x_0$

$$\boxed{? \le Y \le ?}$$

διάστημα πρόβλεψης
forecast interval
prediction interval

Αν ξέρουμε ότι για $X = x_0$ $E(Y) = \mu_0$



① Εκτίμηση του $E(Y)$ όταν $X = x_0$.

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_0$$

σημειακή εκτίμηση της $E(Y|X=x_0)$

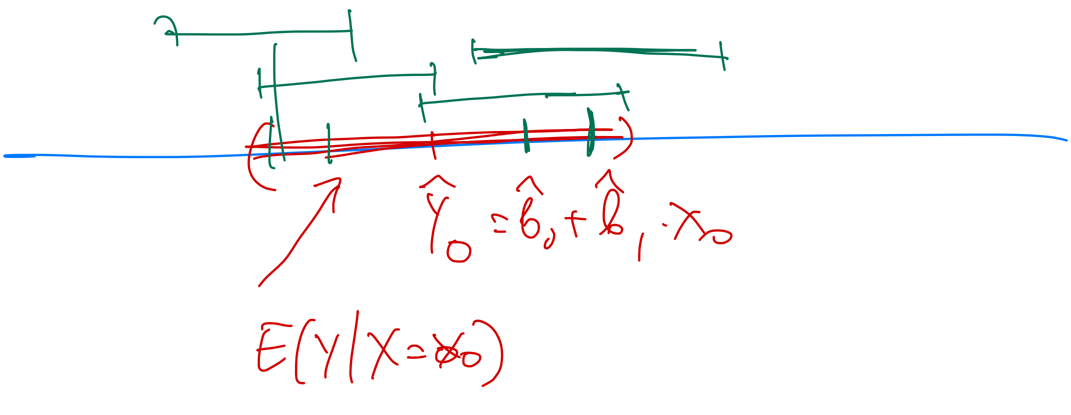
ΔΕ για $E(Y|X=x_0)$ ← διάστημα εμπιστοσύνης

για $X = x_0$. $\hat{Y} = \hat{b}_0 + \hat{b}_1 x_0$ : std. error

$$\text{ΔΕ} = \hat{Y} \pm t_{\alpha/2, dfer}\, s_{\hat{Y}}$$

② Πρόβλεψη { Δεν γνωρίζουμε το $E(Y|X=x_0) = b_0 + b_1 x_0$

Μεμονωμένο άτομο με δική του διακύμανση



Δ Πρόβλεψης $(1-\alpha)\cdot 100\%$

$$\hat{Y} \pm t_{\alpha/2, df_{er}} \cdot \underbrace{S_{f,\hat{Y}_0}}$$

τυπ. σφ. πρόβλεψης

---

Δ Εμ. κ' Προβλεψης έχουν νόημα για τιμές του $X = x_0$ πού είναι μέσα στην περιοχή του X στό δείγμα.

---

Στο R:

---

lm (score~age, data=exercise)

lm (score~age, data=ex2)

| $x$ | $y$ |
|---|---|
| $x_1$ | $y_1$ |
| ⋮ | ⋮ |
| $x_n$ | $y_n$ |

dataset = sample1

δείγμα

$l_1 = lm(y \sim x,\ data = sample1)$

$\Downarrow$

$\hat{b}_0,\ \hat{b}_1,\ MSE,\ R^2 \ldots$

ΔΕ

predict($l_1$, data= sample2, interval="confidence")

| $x$ | $y$ |
|---|---|
| $x'_1$ | ? |
| $x'_2$ | ? |
| ⋮ | ? |
| $x'_k$ | ⋮ |

ΔΠ

predict($l_1$, data=sample2, interval="prediction")

# Μοντέλο Πολλαπλής Παλινδρόμησης

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \varepsilon$$

$b_0, b_1, \ldots, b_k$ : άγνωστα

$\sigma^2$ άγνωστο

**Υποθέσεις**

$$E(\varepsilon) = 0$$

$$Var(\varepsilon) = \sigma^2$$

$$\varepsilon \sim N(0, \sigma^2)$$

$\varepsilon_i, \varepsilon_j$ ανεξάρτητα για διαφορετικές παρατηρήσεις $i, j$

**Δείγμα**

| $j$ | $X_1$ | $X_2$ | $\cdots$ | $X_k$ | $Y$ |
|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{21}$ | | $x_{k1}$ | $y_1$ |
| 2 | $x_{12}$ | $x_{22}$ | | $x_{k2}$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| $n$ | $x_{1n}$ | $x_{2n}$ | | $x_{kn}$ | $y_n$ |

Μέθοδος Ελαχ. Τετραγώνων

$$SSE(b_0, b_1, \ldots, b_k) = \sum_{j=1}^{n} \left( y_j - b_0 - b_1 x_{1j} - b_2 x_{2j} - \cdots - b_k x_{kj} \right)^2$$

$$\min_{b_0, b_1, \ldots, b_k} SSE \implies$$ Εκτιμήτριες ελαχ. τετραγ. $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$

(είναι και εκτιμήτριες μεγ. πιθανοφάνειας)

Εξ. παλινδρόμησης

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \dots + \hat{b}_k x_k \quad \Big\} \text{ response surface}$$

---

Ανάλυση Διασποράς

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})$$

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

---

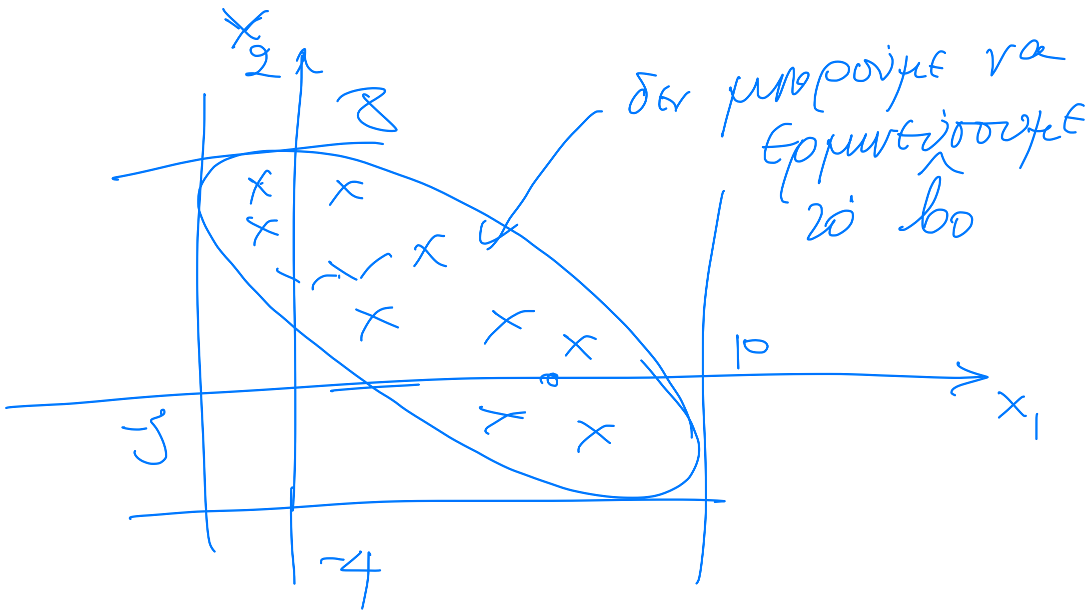
$$SST = SSR + SSE$$

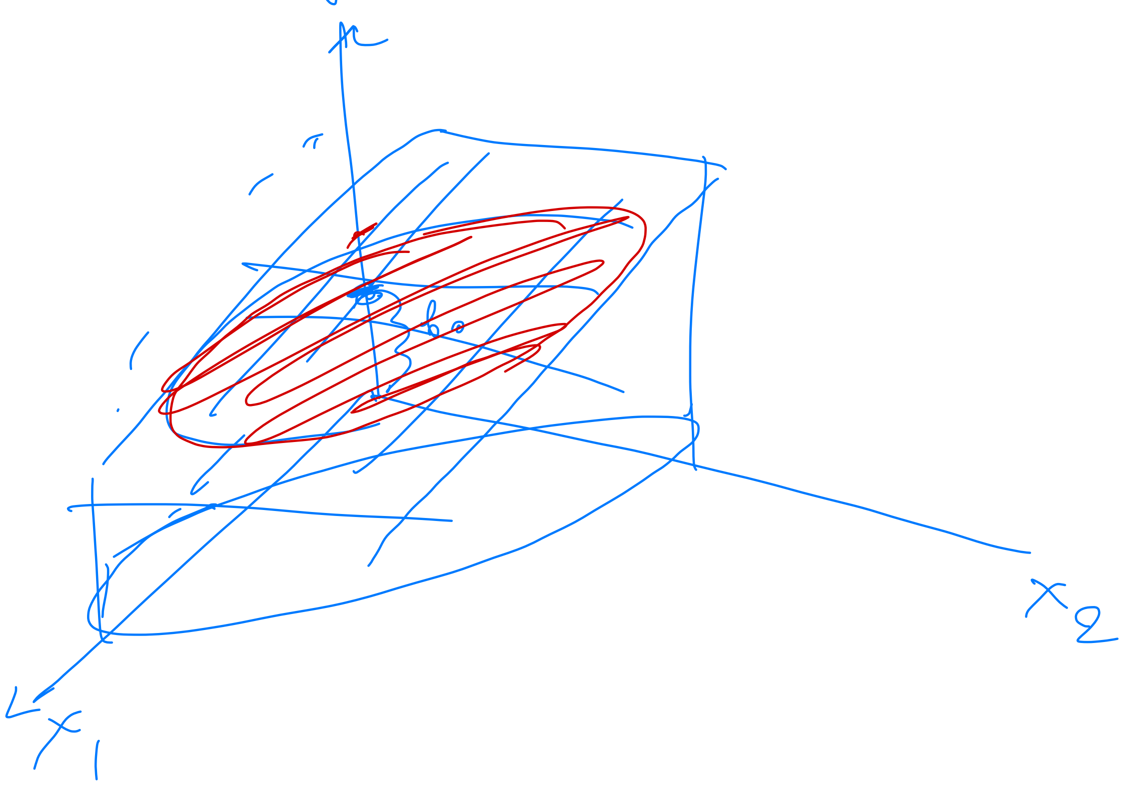
$$R^2 = \frac{SSR}{SST}$$ : % μεταβλ. της Y που εξηγείται από $(x_1, \dots, x_k)$

---

# ① Ερμηνείες

$$\hat{b}_0 = \hat{y} \quad (\text{για } x_1 = 0,\ x_2 = 0,\ \dots,\ x_k = 0)$$



$$EY = b_0 + b_1 x_1 + \cdots + b_k x_k$$

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_k x_k$$

$\hat{b}_1$ = μέση μεταβολή του $Y$ όταν η $x_1$ αυξηθεί κατά 1 κ' όλες οι άλλες μένουν σταθερές

$\hat{b}_2, \ldots, \hat{b}_k$ : αντίστοιχα. (μεμονωμένες επιδράσεις)

$$\hat{b}_1 = \frac{\partial y_1}{\partial x_1}$$

Παρατήρηση

$X_1 \ldots X_k$ δεν είναι απαραίτητο να αντιστοιχούν σε διαφορετικά φυσικά μεγέθη

π.χ. μπορεί $\underbrace{X_2 = X_1^2, \; X_3 = X_1^3 \cdots X_k = X_1^k}_{\text{μοντέλο πολυωνυμικής παλινδρόμησης}}$

ή $\quad x_1 = age \qquad Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$

$\quad x_2 =$ βάρος

Έστω π.χ. $\hat{Y} = \hat{b}_0 + \hat{b}_1 x + \hat{b}_2 x^2$

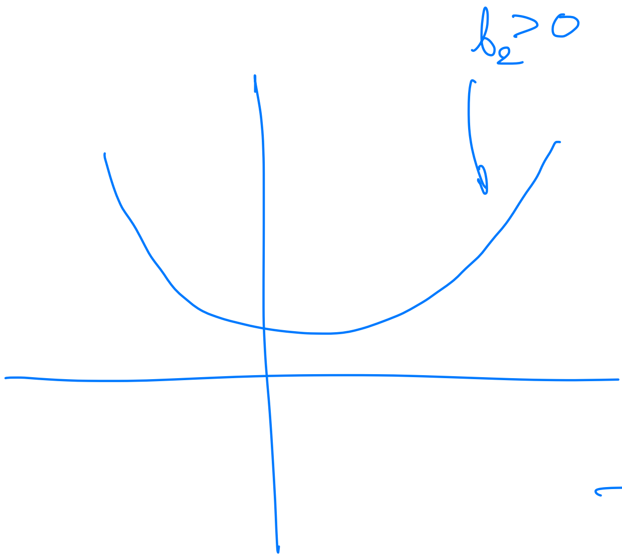
$= \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$

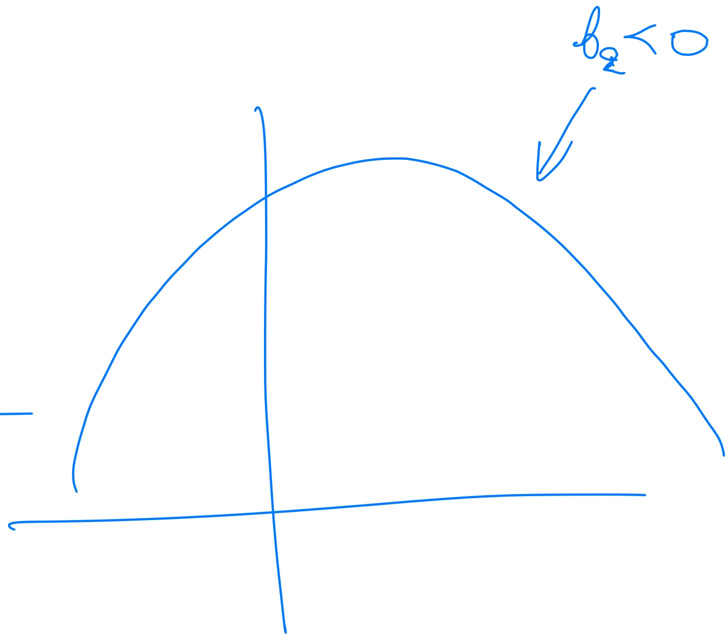
όπου $x_1 = x$

$x_2 = x^2$

$\hat{b}_1 = ?$ δεν μπορούμε να αυξήσουμε το $x_1$ κ' να κρατήσουμε σταθερό το $x_2$

$$\frac{\partial \hat{Y}}{\partial X} = b_1 + 2 b_2 X$$




---

$Y = b_0 + b_1 \log X = b_0 + b_1 x_1$ όπου $x_1 = \log X$

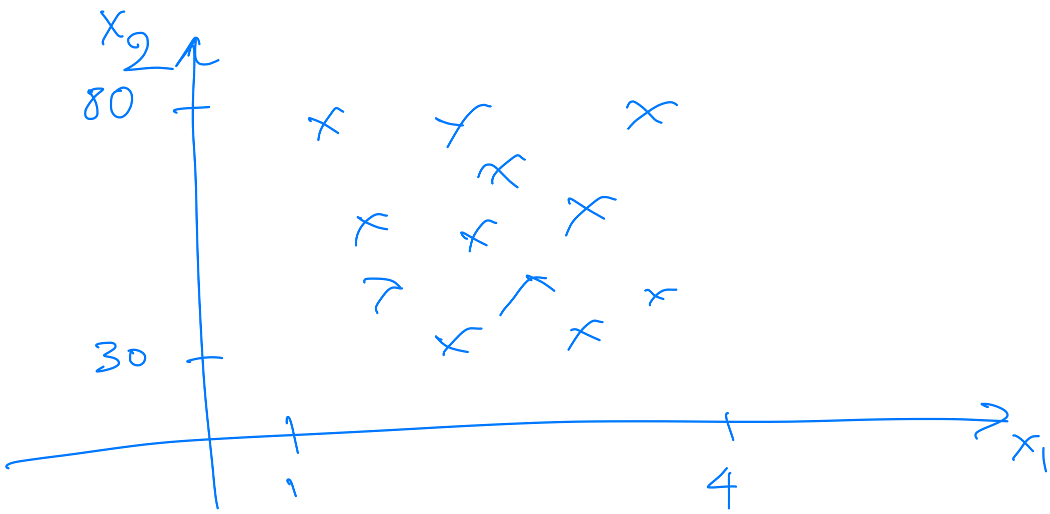
# Παράδειγμα 2

Έστω $X_1$ = age ($1 \leq X \leq 4$ σε έτη)

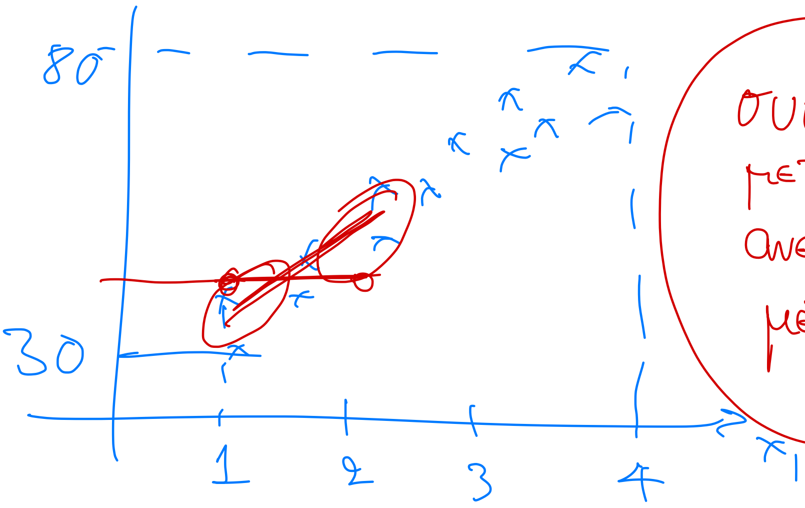
$X_2$ = height (σε cm)

$Y$ = response

$$y = b_0 + b_1 x_1 + \boxed{b_2 x_2}$$



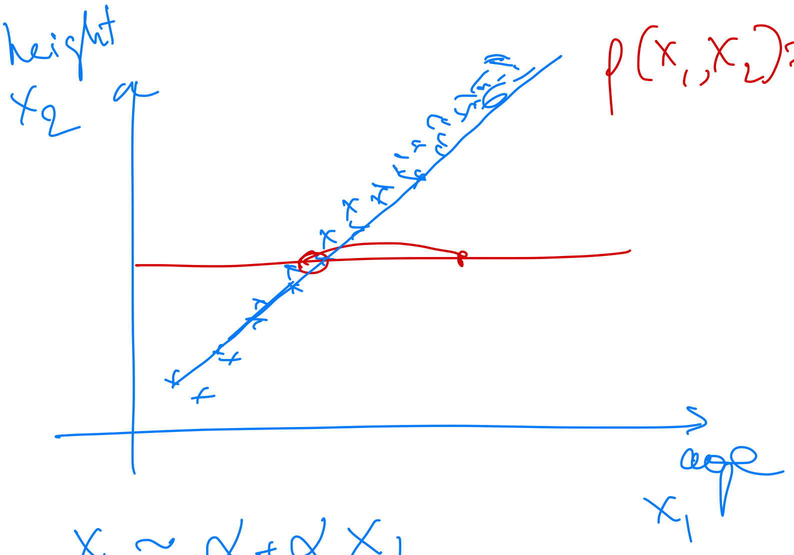

Πιο πιθανό δείγμα

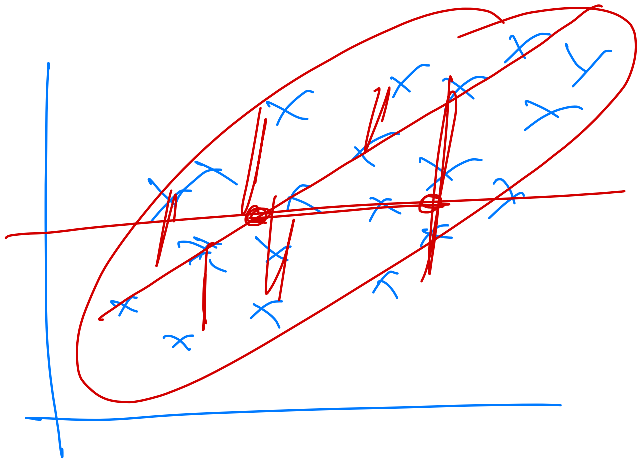



συσχέτιση μεταξύ ανεξάρτητων μεταβλητών

Ερμηνείες των

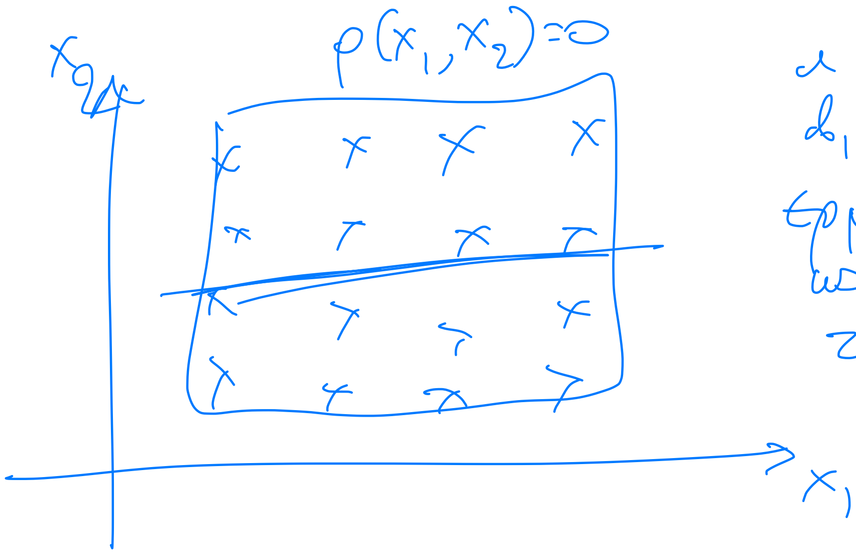
$\rho(x_1, x_2) \approx 1$



$$x_2 \approx \alpha_0 + \alpha_1 x_1$$

$$y = b_0 + b_1 x_1 + b_2 x_2$$
$$= b_0 + b_1 x_1 + b_2(\alpha_0 + \alpha_1 x_1)$$
$$= (b_0 + b_2 a_0) + (b_1 + b_2 a_1) x_1$$




$\hat{b}_1, \hat{b}_2$ έχουν ερμηνεία ως effects των $x_1, x_2$

Έστω $x_1 = age$

$x_2 = height$

$y = βάρος$

---

① Έστω μοντέλο 1 $Y = b_0 + b_1 X_1$ $x_1 = age$.

Έστω $R^2 = 0.70$

② Έστω μοντέλο 2 $Y = b_0 + b_1 x_2$ $x_2 = height$

$R^2 = 0.75$

---

③ Έστω μοντέλο 3 κ' γνωρίζουμε $\boxed{\rho(x_1, x_2) = 0.99}$

$Y = b_0 + b_1 x_1 + b_2 x_2$

$R^2 \approx 0.75$

Mod 1 : $Y = b_0 + b_1\,age$

$R^2 = 0.87$

SSR = 10255910
SSE = 1485592

SST = 11741502

Mod 2 $Y = b_0 + b_1 \cdot extime$ $R^2 = 0.911$

SSR = 10701573
SSE = 1039929

Mod 3 : $Y = b_0 + b_1\,age + b_2\,extime$ $R^2 = 0.92$

SSR = 10255910 +
552338

SSE = 933254

$H_0: b_{age}=0$ $H_1: b_{age} \neq 0$ $p=0.0001$

## Mod 1 (age)

SSR = 10255910

SSE = 1485592

SST = 11741502

SS(age) = 10255910

## Mod 2 (extime)

SSR = 10701573

SSE = 1039929

SST = 11741502

SS(extime) = 10701573

$p=0.32$ $H_0: b_{age}=0$ ; $H_1: b_{age} \neq 0$

## Mod 3 age + extime

SSR = 10808248

SSE = 933254

SST = 11741502

SS(age + extime) = 10808248

SS(age | extime) = SS(age + extime) − SS(extime) = 106675

SS(age) = 10255910

SS(extime + age) = 10808248

SS(extime | age) = SS(extime + age) − SS(age) = 552338

$$Y = b_0 + b_1 \cdot \text{Runtime}$$

$R^2(RT) = 0.90$

$P_{RT} = 0.003$

$$Y = b_0 + b_1\,\text{age} + b_2\,\text{Runtime}$$

$R^2(RT \mid \text{age}) = 0.05$

$P_{RT \mid \text{age}} = 0.45$

An example $P_{RT \mid \text{age}} = 0.03$