

Επανάληψη 30/1/2025

$$R^2 = 0.4721$$

↓
if distance
explains 47%
of total density

$$\rho = 7.5 \cdot 10^5$$

↓
στατ. σημαντική συσχέτιση

$$\hat{b}_0 = 1.21$$

$$\hat{b}_1 = 0.037$$

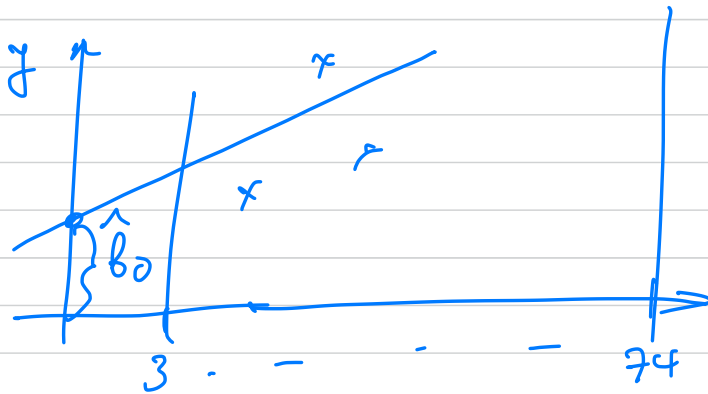
dist

$$\hat{Y} = 1.21 + 0.037 \cdot \text{distance}$$

?
δεν έχει
φυσική
ερμηνεία

↓
για αύξηση απόστασης από ακτί
κάθε 1 km, η μέση πυκνότητα
αυξάνεται κατά 0.0037 gr/km.

b_1 = μεταβολή y ως $\hat{Y}$ ανά μονάδα αύξησης x



3) Για $x = 18 \rightarrow$ πρόβλεψη για $\hat{Y}$
ε' δόσουμε πρόβλεψης

Stat a.

προσδεσμε με γραμμή

μ distance = 18 ϵ' density keró

$\rightarrow$ predict stdf : $\leftarrow$ standard error of prediction

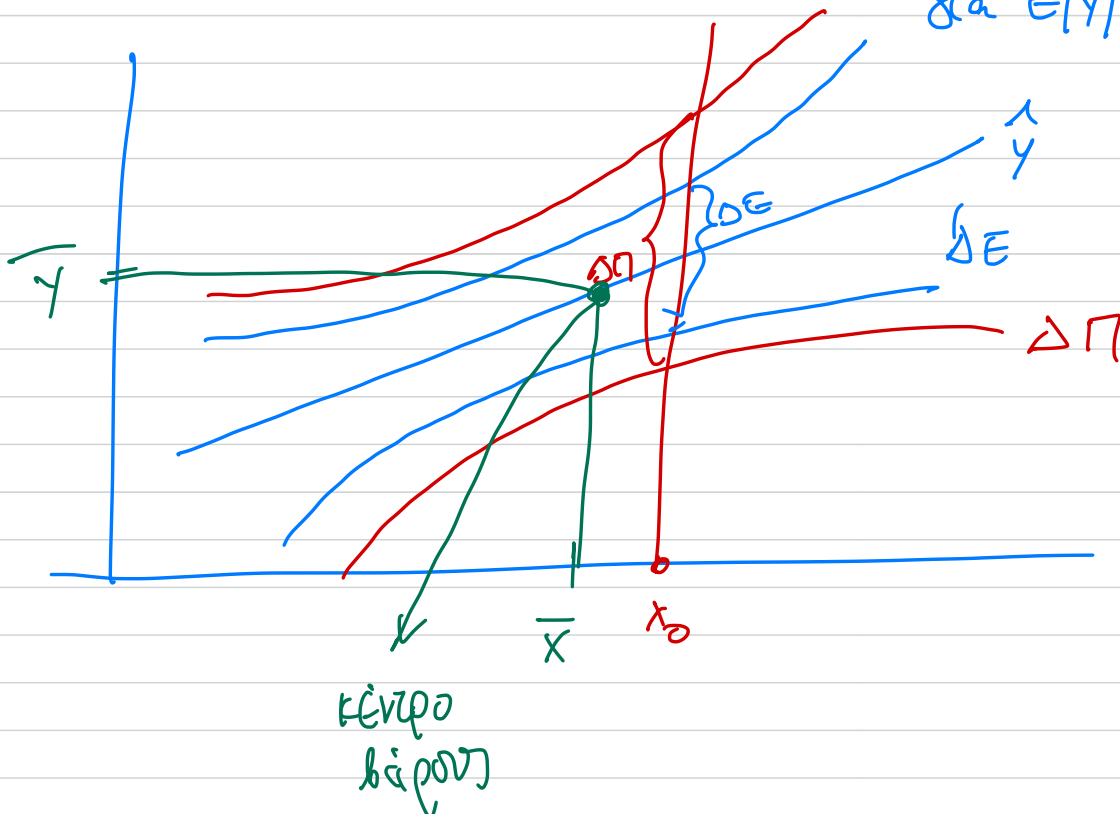
predict $\leftarrow \hat{y}$

$$\text{διασπορά} = \hat{y} + t_{\alpha/2, n-2} \cdot \text{stdf}_{18}$$

R: Συμπεριλάβει το data frame

reef	distance	density
NA	18	NA

$\hat{y} = 1.2797$ $\left\{ \begin{array}{l} (1.072, 1.487) \text{ διασ. 95\% ρεβ. διασποράς για distance = 18} \\ (1.233, 1.326) \text{ διασ. 95\% ρεβ. διασποράς για } E(y/\text{dist}=18) \end{array} \right.$



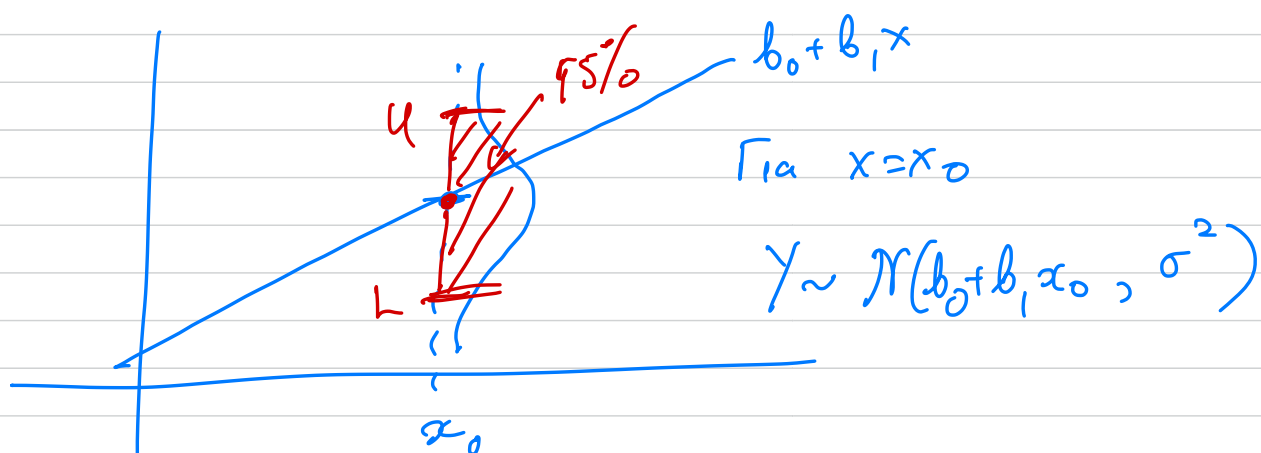
ΔE
 $\mu_{18} E(Y/X=18)$ αγνωστη ποσότητα

Εκτίμηση της $\hat{\mu}_{18} = \hat{Y}(x=18) = \hat{b}_0 + \hat{b}_1 \cdot 18$

standard error της $\hat{\mu}_{18} = \text{stdp}$

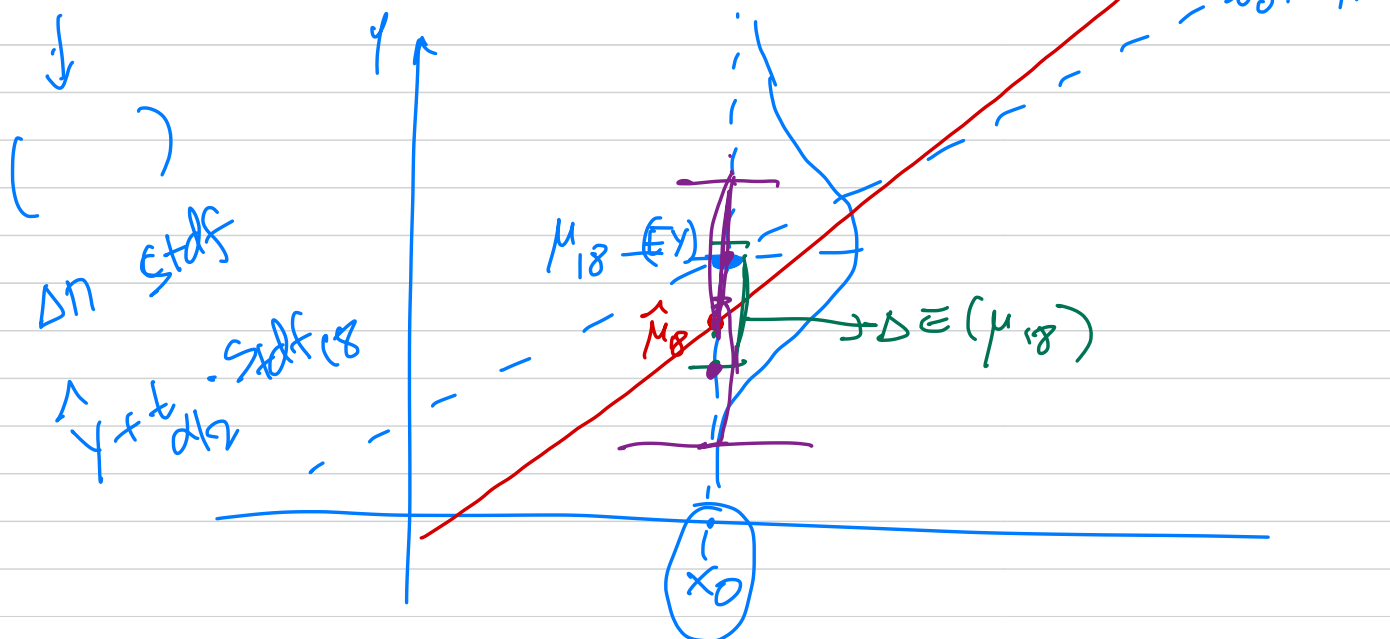
" ΔE " μ_{18} : $\hat{\mu} \pm t_{\alpha/2, n-2} \cdot \text{stdp}$

$\Delta \Pi$ Μια μετασχηματισμένη απάντηση Y για $x=18$



Θέλουμε διάστημα (L, u) : $P(Y \in (L, u)) = 95\%$

Τώρα που δε γνωρίζω τα b_0, b_1



1.4

Outliers

r = Jackknife residuals

Από πίνακα καταλοίπων $k=1$, $n=27$

κρίσιμη τιμή outlier = 3.5

Μια παραζ. θεωρείται outlier αν $|r_{jack}| > 3.5$

Συμφ. $r < -3.5$ ή $r > 3.5$

εδώ όλα r : $-2.45 < r < 1.84$

~~∃~~ outliers

Influential Observations

1) h = leverage : ~~∃~~ influential
κρίσιμη τιμή ≈ 0.35

2) d = Cook's distance

κρίσιμη τιμή πίνακα $\approx \underline{\underline{17.5}}$ για το $d \cdot (n-k-1)$
για το 25d

Μια παρατήρηση είναι influential

αν $25d > 17.5 \Rightarrow d > \frac{17.5}{25} \approx 0.7$

Στα δείγματα όλα $d \leq 0.16 \Rightarrow \text{∄ influential}$

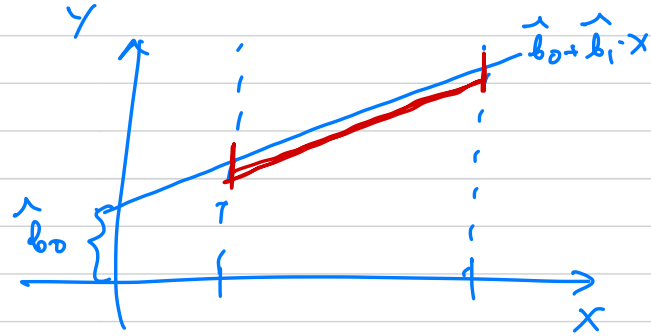
1.5

Shapiro-Wilk test στα rjack:

$p = 0.278 \Rightarrow$ δεν απορρίπτεται
η μηδ. υπόθεση
κανονικότητας

Κεντροποίηση

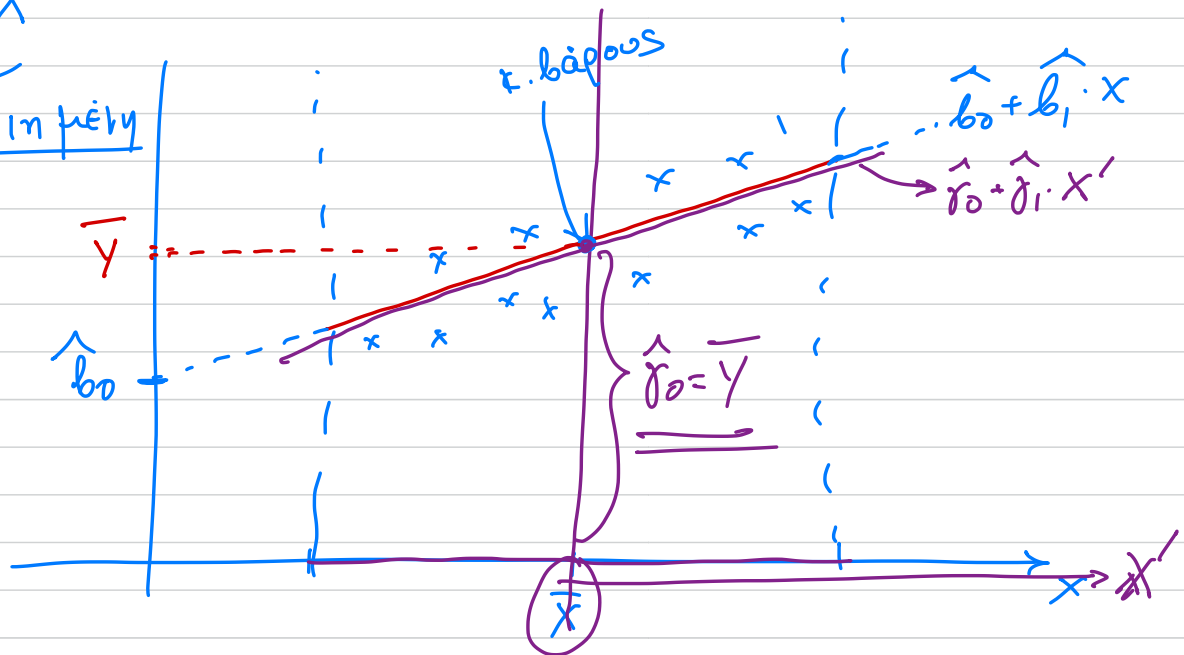
Εστω $Y = b_0 + b_1 X$



Εστω $\bar{X}$ μέση τιμή της X στο δείγμα

$$X' = X - \bar{X}$$

Κεντροποίηση μέγ



Εστω το μορφο

$$Y = \gamma_0 + \gamma_1 X' \\ = \gamma_0 + \gamma_1 (X - \bar{X}) =$$

$$Y = (\gamma_0 - \gamma_1 \bar{X}) + \gamma_1 X$$

$$\Rightarrow \boxed{\hat{b}_0 = \hat{\gamma}_0 - \hat{\gamma}_1 \bar{X}} \quad \text{και} \quad \hat{b}_1 = \hat{\gamma}_1$$

Ξέρουμε ότι αρχικά έχουμε $\hat{b}_0 = \bar{Y} - \hat{b}_1 \cdot \bar{X}$

$$\hat{b}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\Rightarrow \hat{\beta}_0 = \bar{Y}$$

Πρόβλημα 2

Y = undercount (%)

1) Από τα t-test για τα b (δυν. το F-test type III)

λειτουργεί μόνο το perc_min ως perc_min : $p < 0.05$

$$\hat{\beta}_1 = 0.085$$

κρατώντας ως ανεξάρτητες μεταβλητές σταθερές με
αύξηση των ποσοτών μεσοτάξεων κατά 1%

το μέσο ποσοστό υποκαταρκ. αυξάνεται κατά 0.085%

2) Το μοντέλο περιέχει perc_min , crimrate , diffeng , hsgrad

$$\begin{aligned} \hat{Y} = & 0.36 + 0.083 \text{ perc_min} \\ & + 0.0197 \text{ crimrate} \\ & + 0.174 \text{ diffeng} \\ & - 0.049 \text{ hsgrad} \end{aligned}$$

3) Έσω model3 : perc_min , poverty , hsgrad .

(a) Σύγκριση με άλλες μεταβ.

Full model :

$$Y = b_0 + b_1 \text{percmin} + b_2 \text{crimrate} + b_3 \text{poverty} + b_4 \text{diffeng} + b_5 \text{hsgrad} + b_6 \text{housing}.$$

Partial : $Y = b_0 + b_1 \text{percmin} + b_2 \text{poverty} + b_3 \text{hsgrad}$
(nested)

Σύγκριση μέσω F-test

$$H_0 : b_2 = b_4 = b_6 = 0 \quad \text{vs.} \quad H_1 : \text{τουλάχιστον ένα} \neq 0$$

στο full model

Stata : test crimrate diffeng housing

R Τρέχουμε το partial model

`anova(model3, model1)`

$$p\text{-value} = 0.07 > 0.05$$

Δεν μπορεί να απορ. η H_0 :

δεν υπάρχει ένδειξη ότι το model3 είναι
υποβέστερο του model 1

⑥ C_p statistic του partial

Διαπίστωση : θα ξέρουμε τους zινους για
F-test for part k' C_p -mallows

γενικά partial : $Y = b_0 + b_1 X_1 + \dots + b_p X_p$

$$C_p = \frac{SSE(\text{partial})}{MSE(\text{full})} - [n - 2(p+1)]$$

$n = \text{sample size}$
 $p = \# \text{ predictors (excluding the partial)}$

$$n = 66, p = 3$$

$$\text{Anova}(\text{model1}) \Rightarrow MSE(\text{full}) = 2.572$$

$$\text{anova}(\text{model3}) \Rightarrow SSE(\text{partial}) = 170.78$$

$$\Rightarrow C_p = \dots \approx 8.54$$

$$\text{Next best model} \approx \text{full model} \text{ as } C_p \approx p+1 = 4$$

$$\text{Edw } C_p = 8.54 > 4 \Rightarrow \underline{\text{model 3 is the best}}$$

4) housing $\rightarrow$ model $\rightarrow$ p-value = 0.0087 *

housing + pov $\rightarrow$ p-value housing = 0.0845 OK!

Problem 3 $X_N = 1 \{ \text{region} = \text{North} \}$

$$\hat{Y} = 37.87 - 0.017 \cdot \text{elevc} + 5.39 X_N + 0.108 X_N \cdot \text{elevc}$$

$\hat{b}_0 = 37.87 = \hat{Y}$ for $\text{elevc} = 0$ & $X_N = 0$

$\rightarrow$ for $\text{elev} = 927$ & $\text{region} = \text{South}$

$$b_1 = -0.017$$

South :
$$Y = b_0 + b_1 \text{elev} = 37.87 - 0.017 \text{ elev}$$

North :
$$Y = b_0 + b_1 \text{elev} + b_2 + b_3 \cdot \text{elev}$$

$$= (b_0 + b_2) + (b_1 + b_3) \text{elev}.$$

$$= 43.2 + 0.09 \text{ elev}.$$

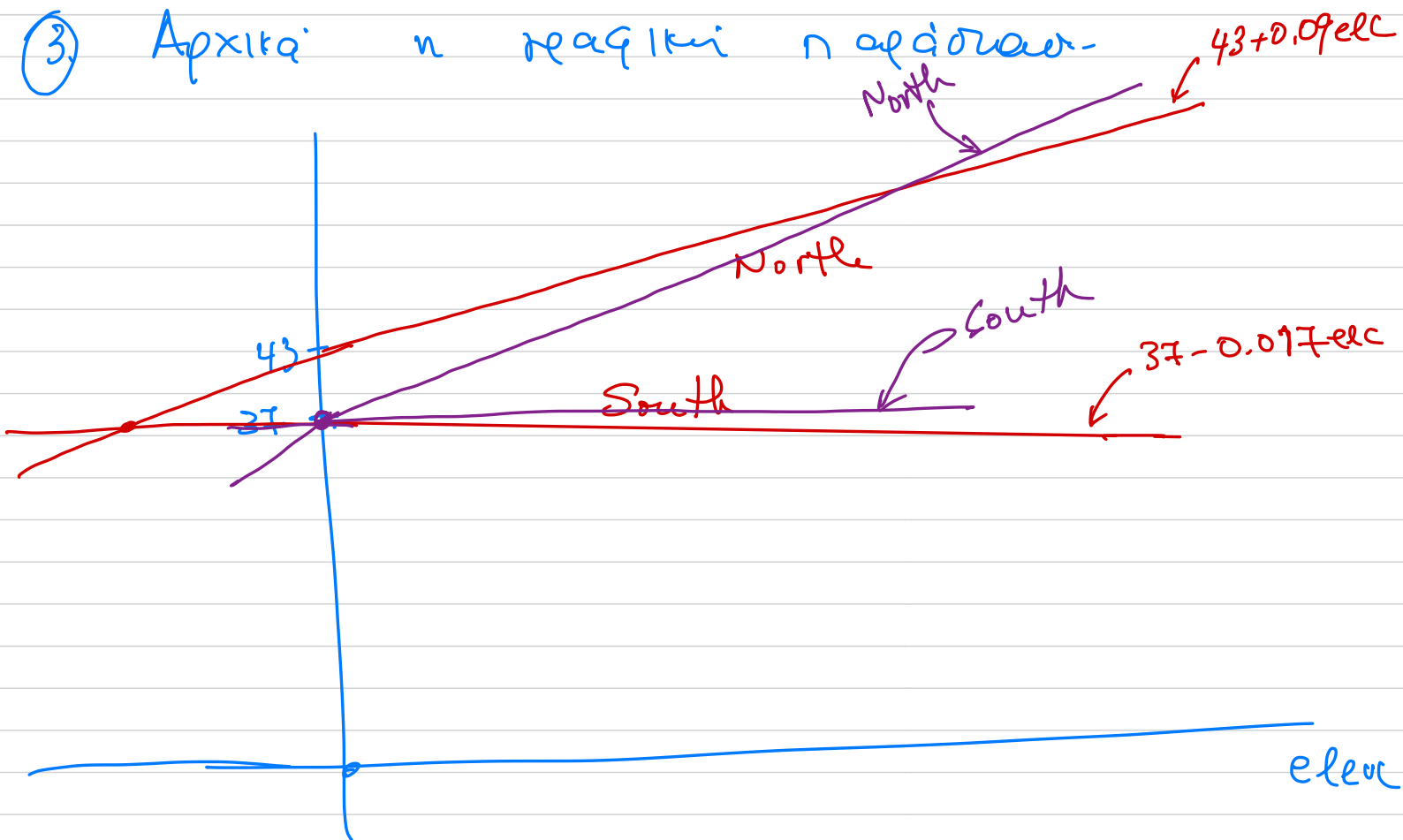
$\hat{b}_1 = -0.017$ = επίπωση υψομέτρου στις νότιες οδούς
 στο τόσο το υψόμετρο όσο και η απόσταση για αύξηση υψομ. κατά 1m
 το ποσοστό βλάβης μειώνεται κατά 0.017%

$\hat{b}_2 = 5.39$ Για elev=0 (δηλ. elev=927)
 η διαφορά στην μέση βλάβη στις
 νότιες οδ. είναι κατά 5.39%
 μεγαλύτερη από όη στις νότιες
 στο μέσο υψόμετρο
 ο νότος θα προσεγγίζει

$\hat{b}_3 = 0.108$: διαφορά στην επίπωση αύξησης υψομέτρου
 κατά 1m μεταξύ οδών, σε North
 κ' South.

για δύο περιοχές σε South κ' North
 η μεταβολή των damage από αύξηση
 υψομέτρου κατά 1m ^{σε North} είναι 0.108%
 μεγαλύτερη από των μεταβολή σε South

③ Αρχικά η γραμμική η απόσταση-



Από το πρώτο στατ. σημαντικό είναι:

Αν αφαιρούσαμε ως μη στατ. σημαντικούς
(ess as main effects)

Εχουμε ως ότι
στην πράξη δεν
αφαιρούμε τα main
effects αν θεωρούμε
τον αλληλεπίδραση

$$\hat{Y} = 37.8 + 0.108 \text{ elevc} \cdot X_N$$

South : $\hat{Y} = 37.8$

North : $\hat{Y} = 37.8 + 0.108 \text{ elevc}$

