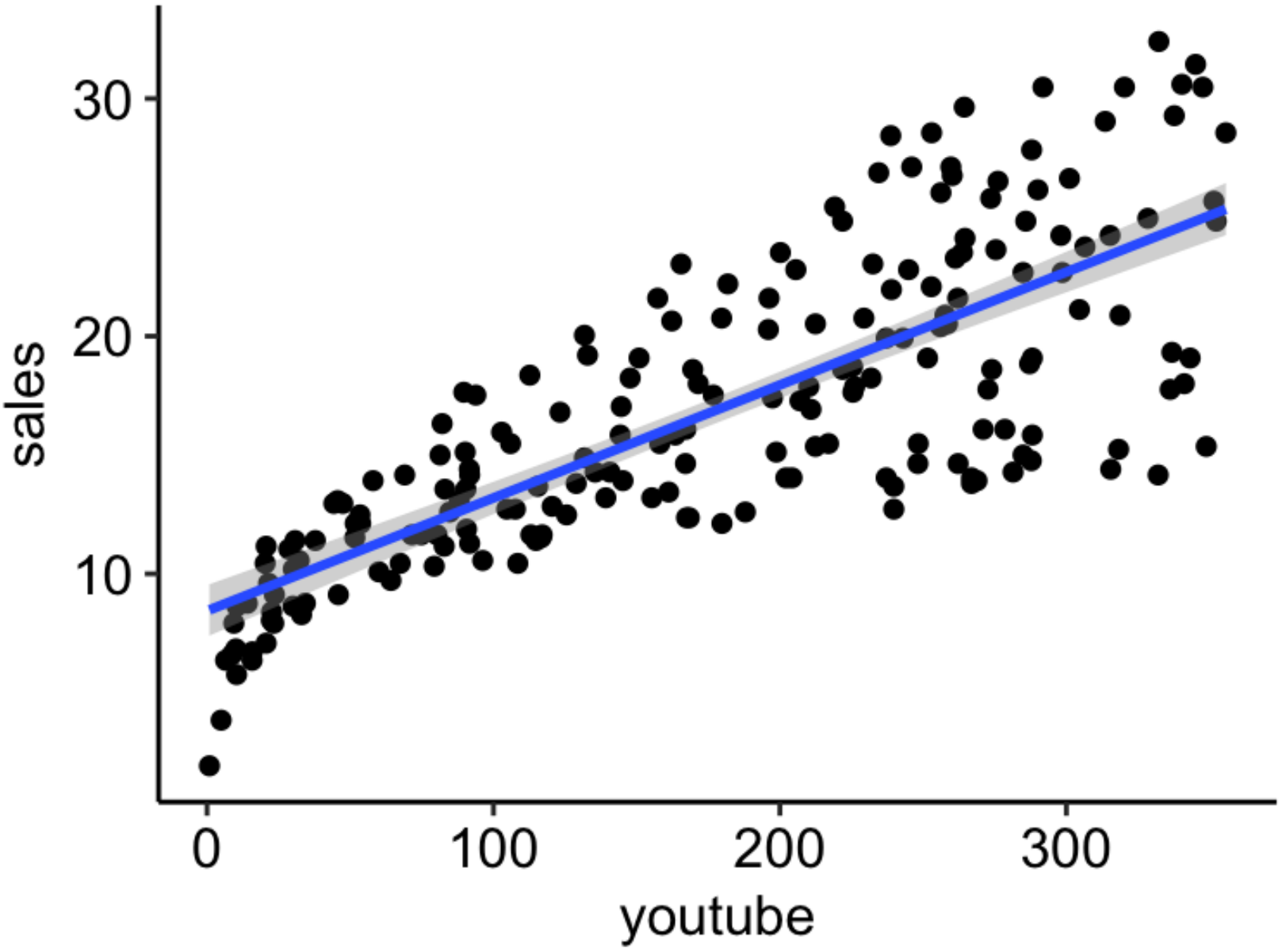
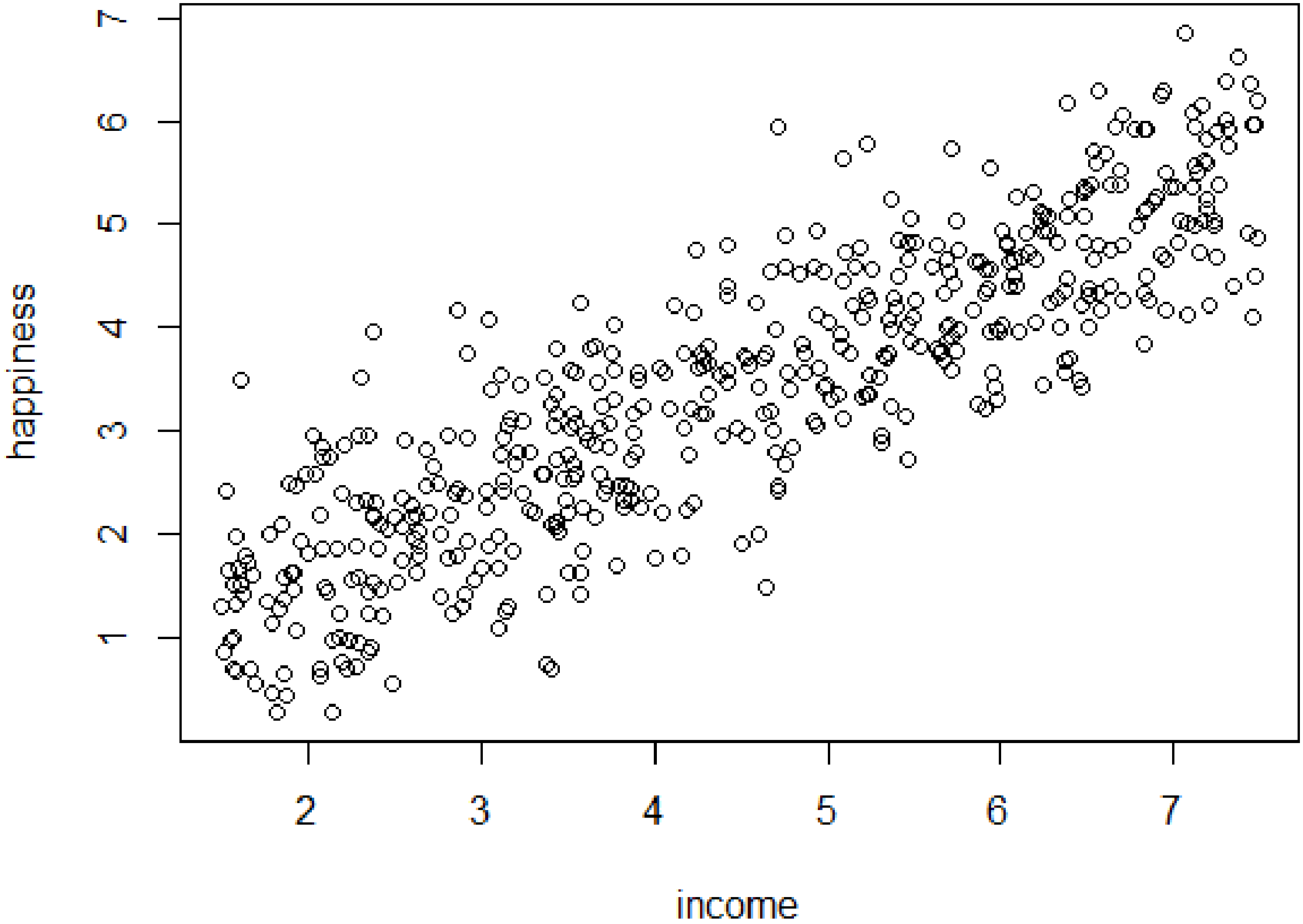
Matematikai statisztika

12. előadás


sales
30
20
10
0
100
200
300
youtube


happiness
1
2
3
4
5
6
7
income
2
3
4
5
6
7

Y közelítése X függvényével

- Gyakori eset, hogy nem ismerjük a számunkra érdekes mennyiség (Y) pontos értékét (pl. holnapi részvény-árfolyam, vízállás, időjárás). Van viszont információnk hozzá kapcsolódó mennyiségről (X, mai értékek).
- Feladat: olyan f_0 megtalálása, amelyre $f_0(X)$ a lehető legjobb közelítése Y-nak.
- Matematikailag: f_0 a megoldása a $\min_f E(Y - f(X))^2$

  szélsőérték-problémának (legkisebb négyzetes becslés).
- Ha az együttes eloszlás ismert (nem teljesen reális, de a megfigyelések alapján közelíthető), akkor megoldható a feladat.

A várható érték optimum-tulajdonsága (ismétlés)

Állítás. A

$$\min_a E(Y-a)^2$$

feladat megoldása $a = E(Y)$.

Bizonyítás. $E(Y-a)^2 = E(Y^2) - 2aE(Y) + a^2$

a szerint deriválva adódik, hogy valóban $E(Y)$ a minimumhely.

A minimum értéke $D^2(Y)$.

Ugyanígy: X tetszőleges értéke esetén $E(Y|X=x)$ adja a minimumot, azaz általánosan $E(Y|X)$ a megoldás.

Optimum a lineáris függvények körében

$$\min_{a,b} E[Y-(aX+b)]^2$$

- Egyszerűbben megoldható
- Nem kell az együttes eloszlás
- A megoldás deriválással:

$$E[Y-(aX+b)]^2 = E(Y^2)+a^2E(X^2)+2abE(X)+$$
$$+b^2-2aE(XY)-2bE(Y)$$

$$\frac{\partial}{\partial a}E[Y-(aX+b)]^2 = 2aE(X^2)+2bE(X)-2E(XY)=0$$

$$\frac{\partial}{\partial b}E[Y-(aX+b)]^2 = 2b+2aE(X)-2E(Y)=0$$

$$aE(X^2) = E(XY)-bE(X)$$

$$b = E(Y)-aE(X)$$

$$aE(X^2) = E(XY)-(E(Y)-aE(X))E(X)$$

$$a = \frac{E(XY)-E(X)E(Y)}{E(X^2)-E^2(X)}$$

$$b = E(Y)-\frac{E(XY)-E(X)E(Y)}{E(X^2)-E^2(X)}E(X)$$

Egyszerű lineáris modell

$$Y_i = aX_i + b + \varepsilon_i, i = 1, \dots, n$$

- X_i a magyarázó változó értéke, ε_i független, azonos eloszlású változók (hibák), $E\varepsilon_i = 0, D^2\varepsilon_i = \sigma^2$, általában feltesszük, hogy normális eloszlásúak.
- a, b, σ^2 a becsülendő együtthatók

Egyenes illesztése az adatokra

Az előzőekhez hasonlóan belátható, hogy a legkisebb négyzetes eltérést

$$\sum_{i=1}^{n}(Y_i - aX_i - b)^2 \to min$$

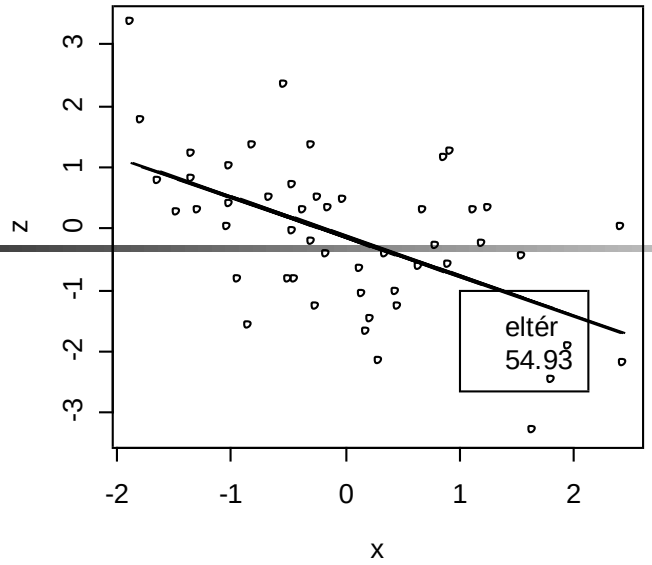
adó egyenes együtthatói:

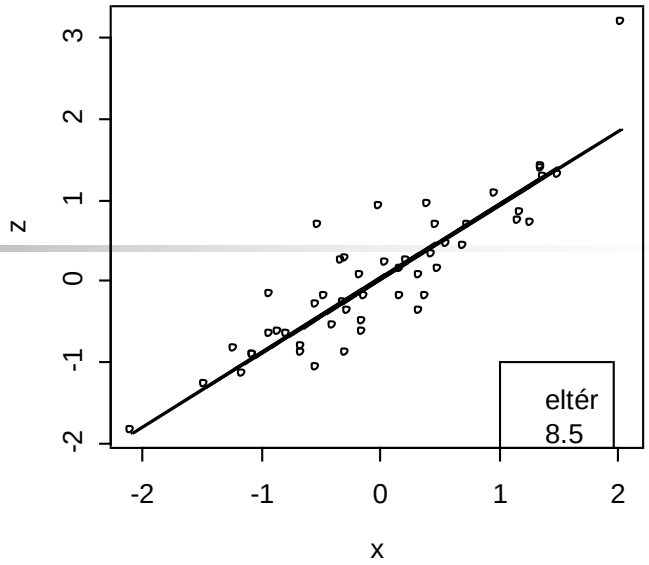
$$a = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

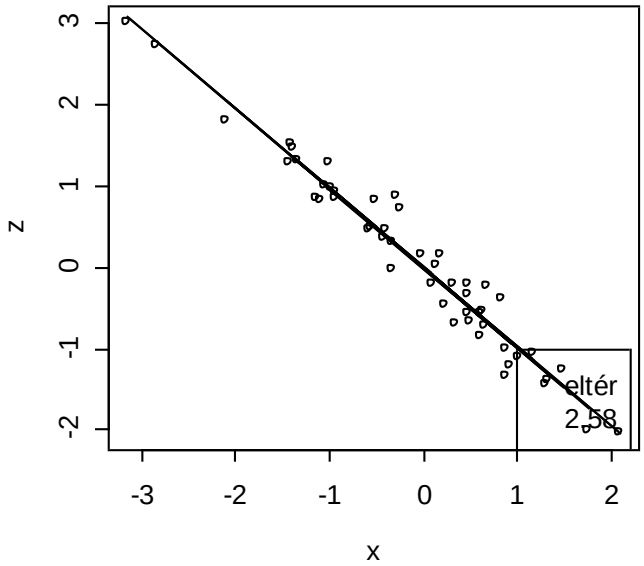
$$b = \bar{Y} - a\bar{X}$$

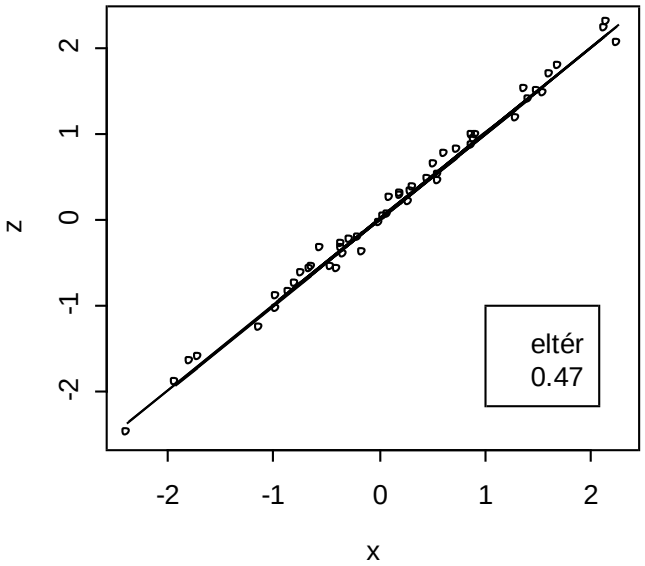
A kapott egyenes tulajdonságai

- Ez a legkisebb négyzetes eltérést adó a lineáris függvények között (a fenti megoldás valóban minimum)
- Elnevezés: regressziós egyenes
- Átmegy az $(\bar{X}, \bar{Y})$ ponton
- a: az X egységnyi változásához tartozó becsült növekmény
- b: az $X = 0$-hoz tartozó becsült érték
- Normális hibák esetén az ML becslés megegyezik ezzel.


z
x
eltér
54.93


z
x
eltér
8.5


z
x
eltér
2.58


z
x
eltér
0.47

Szóródások

- Teljes ingadozás: $\sum_{i=1}^{n}(y_i-\bar{y})^2$
- Reziduális négyzetösszeg: $\sum_{i=1}^{n}(y_i-\hat{a}x_i-\hat{b})^2=\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- A megmagyarázott variabilitás részaránya:

$$R^2=\frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

éppen a tapasztalati korrelációs együttható négyzete

Becslések szórásai és az előrejelzés

$$D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, D(\hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Az x* pontban előrejelzett érték $\hat{a}x^* + \hat{b}$

és ennek szórása $\sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

A szórásbecslésnél σ helyett annak becsült értékét használjuk: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2}{n-2}$

Hipotézisvizsgálat a lineáris regressziónál

Az F-próba speciális esetei:

- H_0:a=0 tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{a}}{\hat{D}(\hat{a})} = \frac{\hat{a}\sqrt{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2}}$$

- Ebből konfidencia intervallum is kapható a-ra

Hipotézisvizsgálat/2

- H_0:b=0 tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{b}\sqrt{n(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \hat{a}x_i - b)^2}\sqrt{\sum_{i=1}^{n}x_i^{\,2}}}$$

- Ebből konfidencia intervallum is kapható b-re, illetve az előrejelzés szórásából az előrejelzett értékre

Többdimenziós lineáris modell

- Több magyarázó változót is bevonhatunk a modellbe:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{k-1} X_{i,k-1} + \varepsilon_i, i = 1, \ldots, n$$

- $Y = X\beta + \varepsilon$

  ahol Y, ε n dimenziós vektorok, X nxk-as mátrix (ismert értékekből), β pedig k dimenziós (ismeretlen) paramétervektor . $E(Y) = X\beta$.

A legkisebb négyzetek módszere

$$\sum_{i=1}^{n} {\varepsilon_i}^2 = (Y - \mathrm{X}\beta)'(Y - \mathrm{X}\beta) \to \min$$

- A megoldás:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

A becslés tulajdonságai

- Torzítatlan
- Kovarianciamátrix:

$$E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \sigma^2(X'X)^{-1}$$

- Ha ε normális eloszlású, akkor a legkisebb négyzetes becslés egyúttal ML becslés is.
- Példák: lineáris regresszió, szórásanalízis.

Hipotézisvizsgálat a lineáris modellben

- A vizsgált hipotézis: $H_0: \beta' H' = 0$

  ahol H rxk-as mátrix (r<k), rang(H)=r.
- A valószínűséghányados próba statisztika:

$$F = \frac{(Y - X\hat{\hat{\beta}})'(Y - X\hat{\hat{\beta}}) - (Y - X\hat{\beta})'(Y - X\hat{\beta})}{(Y - X\hat{\beta})'(Y - X\hat{\beta})}$$

- (n-k)/r F a H_0 esetén F eloszlású (r,n-k) szabadsági fokkal. (Akkor utasítjuk el H_0–t, ha F nagy.)