



Regresszióanalízis

Informatikai Tudományok Doktori Iskola



A regressziószámítás alapproblémája

Regressziószámításkor egy változót egy (vagy több) másik változóval becsülünk.

Y **függőváltozó**

$X_1, X_2, \dots, X_p$ **független változók**

$Y \approx f(X_1, X_2, \dots, X_p)$ becslés $f \in F$

$$E(Y - f(X_1, X_2, \dots, X_p))^2 = \min_{f \in F} E(Y - f(X_1, X_2, \dots, X_p))^2$$



Példák

1. A Duna vízállásának előrejelzése Budapesten
2. A paradicsom beérési idejének becslése
3. Műholdkép alapján a búza terméshozamának becslése
4. Műholdkép alapján a Mars vastartalmának becslése
5. Predikciók, trendek idősoroknál
6. Lineáris közgazdasági modellek



A regressziószámítás alapproblémája

Ha ismerjük az Y és az $X_1, X_2, \dots, X_p$ együttes eloszlását, akkor a probléma elméletileg megoldott:

$$f(X_1, X_2, \dots, X_p) = E(Y | X_1, X_2, \dots, X_p).$$

Gyakorlatban azonban „csak” egy adatmátrix adott:

$$\begin{pmatrix} Y_1 & X_{11} & X_{12} & \dots & X_{1p} \\ Y_2 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$



Feltételes várható érték, folytonos eset I.

Legyenek X és Y folytonos valószínűségi változók!

$F_{X,Y}(x,y)$ és $f_{X,Y}(x,y)$ az együttes eloszlás illetve együttes sűrűségfüggvény!

$$\begin{aligned} P(X < x | y \leq Y < y + \Delta y) &= \frac{P(X < x, y \leq Y < y + \Delta y)}{P(y \leq Y < y + \Delta y)} = \\ &= \frac{F_{X,Y}(x, y + \Delta y) - F_{X,Y}(x, y)}{F_Y(y + \Delta y) - F_Y(y)} = \\ &= \frac{\frac{F_{X,Y}(x, y + \Delta y) - F_{X,Y}(x, y)}{\Delta y}}{\frac{F_Y(y + \Delta y) - F_Y(y)}{\Delta y}} \rightarrow \frac{\partial F_{X,Y}(x, y)}{\partial y} \quad (\Delta y \rightarrow 0). \end{aligned}$$



Feltételes várható érték, folytonos eset II.

$$F_{X|Y}(x|y) \doteq \frac{\partial F_{X,Y}(x, y)}{\partial y}$$

X -nek az Y -ra vonatkozó feltételes eloszlásfüggvénye

$$f_{X|Y}(x|y) \doteq \frac{\partial F_{X|Y}(x|y)}{\partial x} = \frac{\frac{\partial^2 F_{X,Y}(x, y)}{\partial y \partial x}}{f_Y(y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

X -nek az Y -ra vonatkozó feltételes sűrűségfüggvénye



Feltételes várható érték, folytonos eset III.

$$r(y) = \int_{-\infty}^{+\infty} x \cdot f_{X|Y}(x|y) dx = \frac{\int_{-\infty}^{+\infty} x \cdot f_{X,Y}(x,y) dx}{f_Y(y)} \text{ a regressziós görbe}$$

$E(X|Y) = r(Y)$ X -nek Y -ra vonatkozó feltételes várható értéke



A regresszió tulajdonságai

a.) $E(E(X|Y)) = EX$.

b.) $E(h(Y) \cdot X|Y) = h(Y) \cdot E(X|Y)$.

c.) Ha X és Y függetlenek, akkor $E(X|Y) = EX$.

d.) $E(a \cdot X_1 + b \cdot X_2|Y) = a \cdot E(X_1|Y) + b \cdot E(X_2|Y)$.

Az összes függvény közül a regressziós görbével lehet legpontosabban közelíteni!

Legyen $d : \mathbb{R} \rightarrow \mathbb{R}$ tetszőleges függvény.

Ekkor

$$E(X - r(Y))^2 \leq E(X - d(Y))^2, \text{ ahol } E(X|Y) = r(Y).$$



Regresszió normális eloszlás esetén

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\sigma_1}\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2\sigma_1^2(1-\rho^2)} \left[x - \left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(y - \mu_2) \right) \right]^2}$$

$$E(X|Y=y) = \int_{-\infty}^{+\infty} x \cdot f_{X|Y}(x|y) dx = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho(y - \mu_2)$$

Normális komponensek esetén a regressziós összefüggés lineáris!



Elméleti lineáris regresszió

$$a^*X + b^*$$

Y -nak a X -re vonatkozó lineáris regressziója, ha

$$E(Y - a^*X - b^*)^2 = \min_{a,b \in \mathbb{R}} E(Y - aX - b)^2.$$

$$a^* = R(X,Y) \frac{\sigma_Y}{\sigma_X}, \quad b^* = EY - R(X,Y) \frac{\sigma_Y}{\sigma_X} EX.$$



Elméleti lineáris regresszió

Legyen $h(a,b) = E(Y - aX - b)^2$. A lineáris regresszió meghatározásához ezt a kétváltozós függvényt kell minimalizálni. A minimumhely létezésének szükséges feltétele, hogy: $\frac{\partial h}{\partial a} = -2E[(Y - aX - b)X] = 0$, $\frac{\partial h}{\partial b} = -2E[Y - aX - b] = 0$. Innen: $aEX^2 + bEX = EXY$, $aEX + b = EY \Rightarrow b = EY - aEX \Rightarrow aEX^2 + (EY - aEX)EX = EXY \Rightarrow a = R(X,Y) \frac{\sigma_Y}{\sigma_X}$, $b = EY - R(X,Y) \frac{\sigma_Y}{\sigma_X} EX$, $\begin{pmatrix} EX^2 & EX \\ EX & 1 \end{pmatrix}$ pozitív definit, tehát a^*, b^* valóban minimumhely.

Láttuk, hogyha X,Y együttes eloszlása normális, akkor a regresszió lineáris lesz!



A regressziószámítás alapproblémája

$$F = \{f(x_1, x_2, \dots, x_p, a, b, c, \dots \mid a, b, c, \dots \text{ valós paraméterek})\}$$

A függvényhalmazból azt az elemet fogjuk kiválasztani, amelynél:

$$h(a,b,c,\dots) = \sum_{i=1}^n (Y_i - f(X_{1i}, X_{2i}, \dots, X_{pi}, a,b,c,\dots))^2 \rightarrow \min_{a,b,c,\dots}$$

Ez a legkisebb négyzetek módszere!



A regresszióanalízis fajtái

- **Lineáris regresszió** $f(X) = B_0 + B_1 X$

- **Többváltozós lineáris regresszió**

$$f(X_1, X_2, \dots, X_p) = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_p X_p$$

- **Polinomiális regresszió**

$$f(X_1, X_2, \dots, X_p) = B_0 + B_1 X + B_2 X^2 + \dots + B_p X^p$$

$$X_1 = X, X_2 = X^2, \dots, X_p = X^p$$

- **Kétparaméteres (lineárisra visszavezethető) regresszió**

$$\text{pl. } Y = f(X) = B_0 \cdot e^{B_1 X} \Rightarrow \ln Y = B_1 X + \ln B_0$$



A regresszióanalízis fajtái

- **Nemlineáris regressziók két változó között I.**

$$f(X) = B_1 + B_2 \exp(B_3 X) \quad \text{aszimptotikus I.}$$

$$f(X) = B_1 - B_2 \cdot (B_3)^X \quad \text{aszimptotikus II.}$$

$$f(X) = (B_1 + B_2 X)^{-1/B_3} \quad \text{sűrűség}$$

$$f(X) = B_1 \cdot (1 - B_3 \cdot \exp(B_2 X^2)) \quad \text{Gauss}$$

$$f(X) = B_1 \cdot \exp(-B_2 \exp(-B_3 X^2)) \quad \text{Gompertz}$$

$$f(X) = B_1 \cdot \exp(-B_2 / (X + B_3)) \quad \text{Johnson-Schumacher}$$



A regresszióanalízis fajtái

- **Nemlineáris regressziók két változó között II.**

$$f(X) = (B_1 + B_3 X)^{B_2} \quad \text{log-módosított}$$

$$f(X) = B_1 - \ln(1 + B_2 \exp(-B_3 X)) \quad \text{log-logisztikus}$$

$$f(X) = B_1 + B_2 \exp(-B_3 X) \quad \text{Metcherlich}$$

$$f(X) = B_1 \cdot X / (X + B_2) \quad \text{Michaelis Menten}$$

$$f(X) = (B_1 B_2 + B_3 X^{B_4}) / (B_2 + X^{B_4}) \quad \text{Morgan-Merczer-Florin}$$

$$f(X) = B_1 / (1 + B_2 \exp(-B_3 X + B_4 X^2 + B_5 X^3)) \quad \text{Peal-Reed}$$



A regresszióanalízis fajtái

- **Nemlineáris regressziók két változó között III.**

$$f(X) = (B_1 + B_2 X + B_3 X^2 + B_4 X^3) / B_5 X^3 \quad \text{köbök aránya}$$

$$f(X) = (B_1 + B_2 X + B_3 X^2) / B_4 X^2 \quad \text{négyzetek aránya}$$

$$f(X) = B_1 / ((1 + B_3 \cdot \exp(B_2 X))^{(1/B_4)}) \quad \text{Richards}$$

$$f(X) = B_1 / (1 + B_3 \cdot \exp(B_2 X)) \quad \text{Verhulst}$$

$$f(X) = (B_1^{(1-B_4)} \cdot B_2 \exp(-B_3 X))^{1/(1-B_4)} \quad \text{Von Bertalanffy}$$

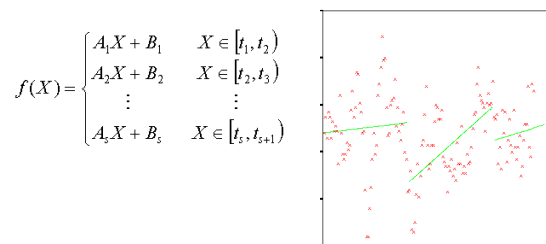
$$f(X) = B_1 - B_2 \exp(-B_3 X^{B_4}) \quad \text{Weibull}$$

$$f(X) = 1 / (B_1 + B_2 X + B_3 X^2) \quad \text{Yield sűrűség}$$



A regresszióanalízis fajtái

- **Szakaszonkénti lineáris regresszió**

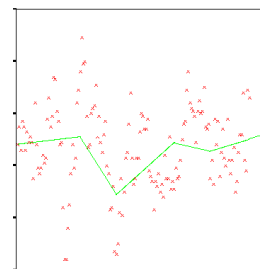


A regresszióanalízis fajtái

- **Poligoniális regresszió**

$$f(X) = \begin{cases} A_1 X + B_1 & X \in [t_1, t_2) \\ A_2 X + B_2 & X \in [t_2, t_3) \\ \vdots & \vdots \\ A_s X + B_s & X \in [t_s, t_{s+1}) \end{cases}$$

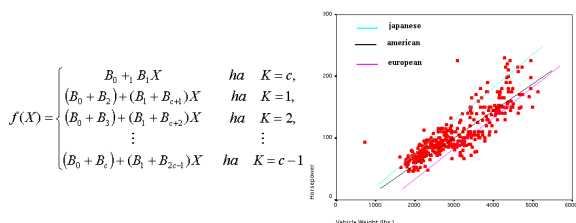
$$A_i t_{i+1} + B_i = A_{i+1} t_{i+1} + B_{i+1}$$





A regresszióanalízis fajtái

- Többváltozós lineáris regresszió kategória-változóval



A regresszióanalízis fajtái

- Logisztikus regresszió

Y dichotóm $Y = \begin{cases} 1, & \text{ha az A esemény bekövetkezik} \\ 0, & \text{ha az A esemény nem következik be} \end{cases}$

A esemény

- A választó fog szavazni
- A páciensnek szívinfarktusa lesz
- Az üzletet meg fogják kötni

$X_1, X_2, \dots, X_p$

ordinális szintű független változók

- eddig hányszor ment el, kor, iskola, jövedelem
- napi cigi, napi pohár, kor, stressz
- ár, mennyiség, piaci forgalom, raktárkészlet



A regresszióanalízis fajtái

- Logisztikus regresszió

$$P(Y=1) = P(A) \approx \frac{1}{1 + e^{-Z}}$$

$$Z = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_p X_p$$

$$\text{ODDS} = \frac{P(A)}{1 - P(A)} \approx e^Z \Rightarrow$$

$$\log(\text{ODDS}) = Z = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_p X_p$$



A regresszióanalízis fajtái

- Logisztikus regresszió

A legnagyobb valószínűség elve

$$L(\epsilon_1, \epsilon_2, \dots, \epsilon_n) = P(Y_1 = \epsilon_1, Y_2 = \epsilon_2, \dots, Y_n = \epsilon_n) =$$

$$= P(Y_1 = \epsilon_1) P(Y_2 = \epsilon_2) \dots P(Y_n = \epsilon_n) \approx$$

$$\approx \frac{1}{1 + e^{-Z_1}} \cdot \frac{1}{1 + e^{-Z_2}} \cdot \dots \cdot \frac{1}{1 + e^{-Z_n}}$$

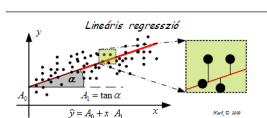
$$\ln L(\epsilon_1, \epsilon_2, \dots, \epsilon_n) = \sum \ln \left(\frac{1}{1 + \exp(B_0 + B_1 X_1 + B_2 X_2 + \dots + B_p X_p)} \right)$$



Lineáris regresszió

A lineáris kapcsolat kitüntetett:

- (1) a legegyszerűbb és leggyakoribb, könnyű a két paramétert értelmezni
- (2) két dimenziós normális eloszlás esetén a kapcsolat nem is lehet más (vagy lineáris vagy egyáltalán nincs)



Lineáris regresszió

Az empirikus lineáris regresszió együtthatóit a legkisebb négyzetek módszerével kaphatjuk meg:

$$h(a, b) = \sum_{i=1}^n (Y_i - aX_i - b)^2 \longrightarrow \min_{a, b}$$

Az empirikus lineáris regresszió együtthatói az elméleti regressziós egyenes együtthatóitól annyiban különböznek, hogy a képletekben az elméleti momentumok helyett a mintából számolt megfelelő empirikus momentumok állnak:

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



Lineáris regresszió

A teljes négyzetösszeg

$$SSTO = Q = \sum_{i=1}^n (y_i - \bar{y})^2$$

A maradékösszeg

$$SSE = Q_{res} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n (y_i - b - a x_i)^2$$

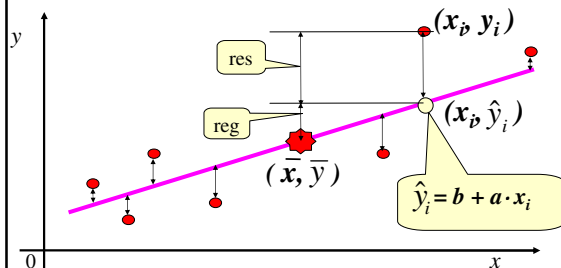
A regressziós összeg

$$SSR = Q_{reg} = \sum_{i=1}^n (\tilde{y}_i - \bar{y})^2$$



A lineáris regresszió

$$Q = Q_{res} + Q_{reg}$$



A lineáris regresszió

A teljes négyzetösszeg felbontása:

$$Q = Q_{res} + Q_{reg}$$

f_{reg} szabadsági foka $n-2$, mert n tagú az összeg, de ezek között két összefüggés van.

f_{res} szabadsági foka mindössze 1, mert az átlag konstans

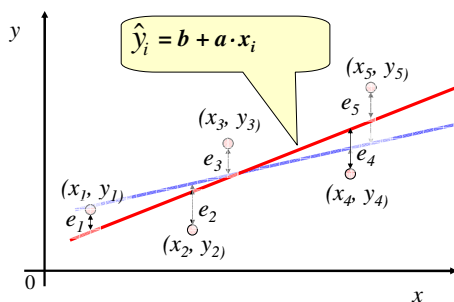
Ha nincs lineáris regresszió, a varianciák hányadosa $(1, n-2)$ szabadsági fokú F eloszlást követ.

$$F = \frac{s_{reg}^2}{s_{res}^2} = \frac{f_{reg}}{f_{res}} = \frac{Q_{reg} (n-2)}{Q_{res}}$$



A lineáris regresszió

A legkisebb négyzetek módszere alapelve:



A lineáris regresszió

Megjegyzések:

1.

Ha X és Y együttes eloszlása normális, akkor az elméleti regressziós egyenes a meredekségére konfidenciaintervallum szerkeszthető, mivel ilyenkor az $\frac{a^* - a}{\frac{s_y}{s_x} \sqrt{1 - R_n^2}} \sqrt{n-2}$ statisztika $n-2$ szabadságfokú Student-eloszlást követ.

2.

A normális esetben a korrelálatlanság és a függetlenség azonos tulajdonságok. Tehát, ha X és Y korrelációs együtthatója 0, akkor $a = 0$, azaz

$$\frac{a^*}{\frac{s_y}{s_x} \sqrt{1 - R_n^2}} \sqrt{n-2} = \frac{R_n}{\sqrt{1 - R_n^2}} \sqrt{n-2} \in t_{n-2}.$$

A függetlenséget megfogalmazó nullhipotézisről tehát ilyenkor t-próbával dönthetünk.



A lineáris regresszió

Tervezett (determinisztikus) megfigyelés

Főleg műszaki alkalmazásokban gyakori, hogy a méréseket Y -ra előírt x beállításoknál végzik el, és így keresik az ismeretlen $Y \sim f(x)$ függvénykapcsolatot. A modell ilyenkor az, hogy

$Y = f(x) + \varepsilon$, ahol ε a mérési hibát jelentő valószínűségi

változó, melyre $E \varepsilon_i = 0$ és $\sigma^2 \varepsilon_i = D^2 < \infty$. Tegyük fel, hogy adottak az $x_1, x_2, \dots, x_n \in \mathbb{R}$ beállítások mellett elvégzett $Y_1, Y_2, \dots, Y_n$ mérési eredmények. Mivel a mérések a véletlentől is függetlek, feltesszük, hogy $Y_i = ax_i + b + \varepsilon_i$ ($i = 1, \dots, n$), ahol ε_i teljesen függetlenek és $E \varepsilon_i = 0$, $\sigma^2 \varepsilon_i = D^2 < \infty$. A keresett a, b regressziós együtthatókat a legkisebb négyzetek módszerével a

$$h(a, b) = \frac{1}{n} \sum_{i=1}^n (Y_i - (ax_i + b))^2 = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$$

négyzetes eltérés átlagának minimalizálásával oldjuk meg.



Gauss-Markov-tétel

Ha $Y_i = ax_i + b + \varepsilon_i$ ($i = 1, 2, \dots, n$), ahol az ε_i teljesen független valószínűségi változók, és $E\varepsilon_i = 0$, $\sigma^2\varepsilon_i = \sigma^2$, akkor az a, b együtthatók legkisebb négyzetek módszerével kapott becslései torzítatlanok, és az összes lineáris becslés közül minimális szórással rendelkeznek.

Megjegyzés: A legkisebb négyzetek módszere a legjobb torzítatlan becslést adja, ami angolul: *best linear unbiased estimation* = BLUE.

Ha $Y_i = ax_i + b + \varepsilon_i$ ($i = 1, 2, \dots, n$), ahol az $\varepsilon_i \in N(0, \sigma^2)$ teljesen független valószínűségi változók ($\Rightarrow Y_i \in N(ax_i + b, \sigma)$ és teljesen függetlenek), akkor az előbbiek mellett még az is állítható, hogy a^* és b^* az a, b paramétereknek maximum-likelihood becslései is.



Lineárisra visszavezethető kétparaméteres regresszió

$$\mathfrak{S} = \{f(x; a, b)\} \quad E(Y - f(X; a^*, b^*))^2 = \min_{a, b} E(Y - f(X; a, b))^2$$

Amennyiben találhatók olyan alkalmas f_{1, k_2} függvények, amivel

a probléma linearizálható: $y = f(x; a, b) \quad (y) = k_1(a, b) \cdot h(x) + k_2(a, b)$

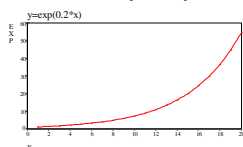
A trükkkel nem az eredeti minimalizálási feladat megoldását kapjuk meg, csak attól nem túl messze eső közelítéseket!



Lineárisra visszavezethető kétparaméteres regresszió

exponenciális függvénykapcsolat: $y = a \cdot e^{bx}$

$$y^* = \ln y = b \cdot \ln x + \ln a = k_1 \cdot x + k_2$$



„growth” függvény: $y = \exp(a + bx)$ $y^* = \ln y = ax + b$

„compound” függvény: $y = a \cdot b^x$ $y^* = \ln y = \ln a + x \cdot \ln b$

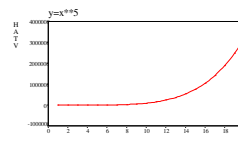


Lineárisra visszavezethető kétparaméteres regresszió

hatványfüggvény: $y = a \cdot x^b$

$$y^* = \ln y = b \cdot \ln x + \ln a = k_1 \cdot x^* + k_2$$

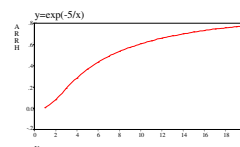
$$\Rightarrow b = k_1, \quad a = e^{k_2}$$



Arrhenius: $y = a \cdot e^{-\frac{b}{x}}$

$$y^* = \ln y = -b \cdot \frac{1}{x} + \ln a = k_1 \cdot x^* + k_2$$

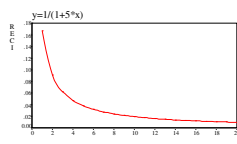
$$b = -k_1, \quad a = e^{k_2}$$



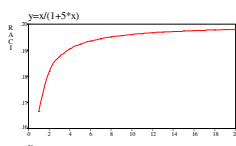
Lineárisra visszavezethető kétparaméteres regresszió

reciprok: $y = \frac{1}{a + b \cdot x}$

$$y^* = \frac{1}{y} = a \cdot x + b$$



racionalis: $y = \frac{a \cdot x}{1 + b \cdot x}$



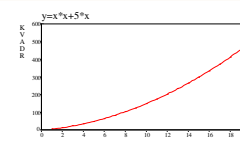
$$y^* = \frac{1}{y} = \frac{1}{a} \cdot \frac{1}{x} + \frac{b}{a} = k_1 \cdot x^* + k_2 \Rightarrow a = \frac{1}{k_1}, \quad b = \frac{k_2}{k_1}$$



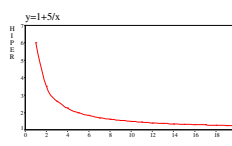
Lineárisra visszavezethető kétparaméteres regresszió

homogén kvadratikusan: $y = a \cdot x^2 + b \cdot x$

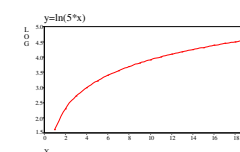
$$y^* = \frac{y}{x} = a \cdot x + b$$



hiperbolikus: $y = a + \frac{b}{x}$

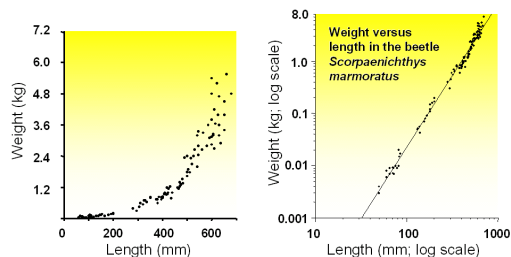


logaritmus: $y = a \cdot \ln(b \cdot x) = a \cdot \ln b + a \cdot \ln x$





Linearizálás, pl.



Polinomiális regresszió

Amikor az $F = \{p_n(x) = a_0 + a_1x + \dots + a_mx^m\}$ függvényosztály a legfeljebb m -edrendű polinomosztály, a $\min_{Y_j \in F} \mathbf{E}(Y - f(X))^2$ minimumfeladat megoldását *polinomiális regresszió* illesztésnek nevezzük.

Az elméleti polinomiális regresszió

A polinomiális regressziós feladatot többváltozós lineáris regresszióval oldhatjuk meg, a prediktor változók ilyenkor az X változó hatványai: $X_i = X^i$!

lineáris egyenletrendszer megoldásával kaphatjuk meg. Ennek mindig van megoldása, hiszen az együttthatómátrix szimmetrikus és pozitív szemidefinit.



Polinomiális regresszió

Bizonyítás: A feladatot a

$$h(a_0, a_1, \dots, a_m) = \mathbf{E}(Y - (a_0 + a_1X + \dots + a_mX^m))^2$$

$m+1$ változós függvény minimumhelyének megkeresésével oldhatjuk meg:

$$\frac{\partial h(a_0, a_1, \dots, a_m)}{\partial a_i} = -2\mathbf{E}[(Y - (a_0 + a_1X + \dots + a_mX^m))X^i] = 0 \quad (i = 0, 1, 2, \dots, m) \Rightarrow \sum_{j=0}^m a_j \mathbf{E}X^{i+j} = \mathbf{E}YX^i \Rightarrow \text{következik az állítás.}$$



Polinomiális regresszió

A tapasztalati polinomiális görbe együtthatóinak meghatározását az $(X_1, Y_1)^T, (X_2, Y_2)^T, \dots, (X_n, Y_n)^T$ statisztikai minta segítségével az

$$\begin{pmatrix} 1 & \frac{1}{n} \sum_{j=1}^n X_j & \dots & \frac{1}{n} \sum_{j=1}^n X_j^m \\ \frac{1}{n} \sum_{j=1}^n X_j & \frac{1}{n} \sum_{j=1}^n X_j^2 & \dots & \frac{1}{n} \sum_{j=1}^n X_j^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} \sum_{j=1}^n X_j^i & \frac{1}{n} \sum_{j=1}^n X_j^{i+1} & \dots & \frac{1}{n} \sum_{j=1}^n X_j^{i+m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} \sum_{j=1}^n X_j^m & \frac{1}{n} \sum_{j=1}^n X_j^{m+1} & \dots & \frac{1}{n} \sum_{j=1}^n X_j^{2m} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_i \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{j=1}^n Y_j \\ \frac{1}{n} \sum_{j=1}^n Y_j X_j \\ \vdots \\ \frac{1}{n} \sum_{j=1}^n Y_j X_j^i \\ \vdots \\ \frac{1}{n} \sum_{j=1}^n Y_j X_j^m \end{pmatrix}$$

lineáris egyenletrendszer megoldásából kapjuk.



A regresszió közelítése Nadaraja módszerével

Az X és Y változók között a „tökéletes” függvénykapcsolatot az $r(x) = \mathbf{E}(Y|X=x)$ regressziós görbe adja meg.

Nadaraja nemparaméteres módszere a sűrűségfüggvény Parzen-Rosenblatt becslését használja.

A sűrűségfüggvény becslését felhasználva a

$$\mathbf{E}(Y|X=x) = r(x) = \int_{-\infty}^{+\infty} y \cdot \frac{f_{X,Y}(x,y)}{f_X(x)} dy$$

regressziós görbét „közvetlenül” becsli.



A regresszió közelítése Nadaraja módszerével

Tétel: Legyen az $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ minta együttes sűrűségfüggvénye $f(x,y)$. Legyen továbbá $k(x)$ olyan páros sűrűségfüggvény, amelyre igazak a következők:

- (i) $k(x)$ korlátos függvény
- (ii) $|x \cdot k(x)| \rightarrow 0$ ha $|x| \rightarrow \infty$
- (iii) $k(x)$ második momentuma véges

Legyen a $h_n > 0$ számsorozat olyan, hogy

- (iv) h_n nullsorozat
- (v) $n \cdot h_n \rightarrow \infty$

Akkor az

$$r_n(x) = \frac{\sum_{i=1}^n Y_i \cdot k\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n k\left(\frac{x - X_i}{h_n}\right)}$$

az $r(x)$ regressziós görbe konzisztens becslése.

Pl. $k(x) = \varphi(x)$ és $h_n = n^{-1/3}$ jó választás,

$\forall \varepsilon > 0: \mathbf{P}[|r_n(x) - r(x)| > \varepsilon] \rightarrow 0$ ($n \rightarrow \infty, \forall x \in \mathbb{R}$)



A regresszió közelítése Nadaraja módszerével

Mivel minden esetben bár nagy, de mégiscsak véges mintával végezzük a becslést, a h_n sorozat megadása helyett a

$$r_n(x) = \frac{\sum_{i=1}^n y_i \cdot k\left(\frac{x-x_i}{h}\right)}{k\left(\frac{x-x_i}{h}\right)}$$

kifejezésben a h paraméterrel minimalizálunk.



Egy példa az alkalmazásra

Egy meteorológiai mérőballon segítségével különböző magasságokban megmérték a levegő napi ózon szintjét.

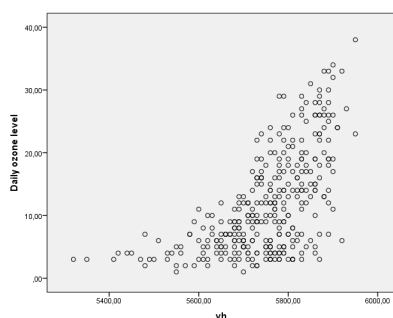
Az összesen $n=330$ napon mértek:

	vh	ozon	alt	deg	vis	temp	day
1	5710.00	3.00	2693.00	-25.00	250.00	40.00	3.00
2	5700.00	5.00	590.00	-24.00	100.00	45.00	4.00
3	5760.00	5.00	1450.00	25.00	60.00	54.00	5.00
4	5720.00	6.00	1568.00	15.00	60.00	35.00	6.00
5	5790.00	4.00	2631.00	-33.00	100.00	45.00	7.00
6	5790.00	4.00	1554.00	-26.00	250.00	55.00	8.00
7	5700.00	5.00	3093.00	23.00	120.00	41.00	9.00
8	5700.00	7.00	2654.00	-2.00	120.00	44.00	10.00
9	5770.00	4.00	5000.00	-19.00	120.00	54.00	11.00
10	5720.00	6.00	111.00	9.00	150.00	51.00	12.00
11	5760.00	5.00	492.00	-44.00	40.00	51.00	13.00
12	5780.00	4.00	5000.00	-44.00	200.00	54.00	14.00
13	5830.00	4.00	1249.00	-53.00	250.00	59.00	15.00
14	5870.00	7.00	5000.00	-67.00	200.00	51.00	16.00
15	5840.00	5.00	5000.00	40.00	200.00	64.00	17.00
16	5790.00	9.00	639.00	1.00	150.00	67.00	18.00
17	5680.00	4.00	393.00	-68.00	10.00	52.00	19.00
18	5720.00	3.00	5000.00	-66.00	140.00	54.00	20.00
19	5760.00	4.00	5000.00	-58.00	250.00	54.00	21.00
20	5730.00	4.00	5000.00	-26.00	200.00	58.00	22.00



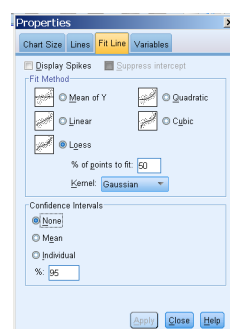
Egy példa az alkalmazásra

A két változó szóródásábrája:



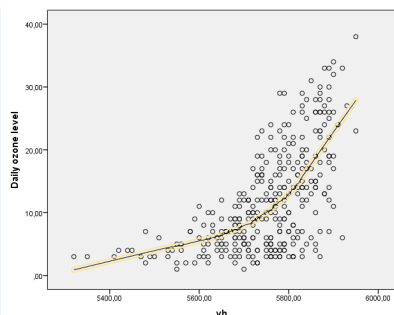
Egy példa az alkalmazásra

Gaussi magfüggvényt használva:



Egy példa az alkalmazásra

A regresszió Nadaraja becslése:



A többváltozós lineáris regresszió

A független változók azon lineáris kombinációját keressük, amelynél a függőváltozót legkisebb négyzetes hibával tudjuk közelíteni:

$$Y_i = \tilde{Y}_i + \varepsilon_i = b_0 + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + \varepsilon_i$$

$$\begin{aligned} \sum_{i=1}^n (Y_i - \tilde{Y}_i)^2 &= \|\varepsilon\|^2 = \sum_{i=1}^n \varepsilon_i^2 = (\underline{Y} - \underline{X}\underline{b})^T (\underline{Y} - \underline{X}\underline{b}) = \\ &= \underline{Y}^T \underline{Y} - \underline{b}^T \underline{X}^T \underline{Y} - \underline{Y}^T \underline{X} \underline{b} + \underline{b}^T \underline{X}^T \underline{X} \underline{b} = \\ &= \underline{Y}^T \underline{Y} - 2 \underline{Y}^T \underline{X} \underline{b} + \underline{b}^T \underline{X}^T \underline{X} \underline{b} = Q(\underline{b}) \rightarrow \min_{\underline{b}} \end{aligned}$$



A többváltozós lineáris regresszió

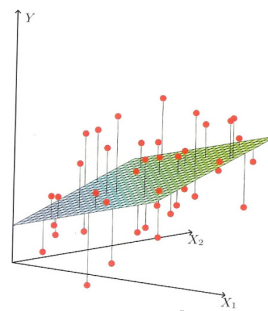
Az együtthatók meghatározása a legkisebb négyzetek módszerével:

$$\frac{\partial Q}{\partial \underline{b}} = -2 \underline{X}^T \underline{Y} + 2 \underline{X}^T \underline{X} \underline{b} = 0$$

$$\underline{X}^T \underline{X} \underline{b} = \underline{X}^T \underline{Y} \Rightarrow \underline{b} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$$



A többváltozós lineáris regresszió



A többváltozós lineáris regresszió

Szórásanalízis (ANOVA) a modell érvényességének eldöntésére

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$\frac{SSR / (k-1)}{SSE / (n-k)} \in F_{k-1, n-k}$$

A nullhipotézis az, hogy a független változók mindegyike 0, vagyis egyik prediktor változó sem magyarázza a célváltozót!

F-próbával dönthetünk a nullhipotézisről.



A többváltozós lineáris regresszió

Béta-együtthatók

$$BETA_i = b_i \cdot \frac{S_y}{S_i} \quad (i = 1, 2, \dots, k)$$

b_i az i -edik regressziós

S_i az i -edik változó

S_y

a célváltozó standard szórása.

A béta-együtthatók egyfajta szempontból minősítik a változók fontosságát a lineáris összefüggésben. Ha egy változónak nagy az együtthatója abszolút értékben, akkor fontos, ha kicsi, kevésbé fontos.



A többváltozós lineáris regresszió

R^2 (coefficient of determination) meghatározottsági együttható

$$SSR = Q_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SSTO = Q = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}, \quad 0 \leq R^2$$

Ha csak egy magyarázó változó van, akkor R^2 éppen a korrelációs együttható négyzetek!

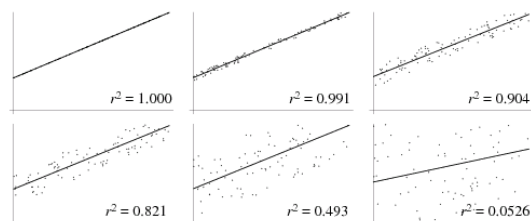
$$R = \pm \sqrt{\frac{SSR}{SSTO}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad -1 \leq R \leq 1.$$

Megmutatja, hogy a lineáris regresszióval a célváltozó variációjának mekkora hányadát lehet magyarázni



A többváltozós lineáris regresszió

Az R^2 érték megmutatja a lineáris kapcsolat mértékét





A többváltozós lineáris regresszió

Korrigált (adjusztált) meghatározottsági mutató

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1} = 1 - \frac{SSE / (n-p-1)}{SSTO / (n-1)}$$

p a független változók száma

A korrekció azért szükséges, mert újabb változók bevonásával R^2 automatikusan nő, és túl optimista képet mutat a modell illeszkedéséről. Az adjusztált változatban „büntetjük” a túl sok változó bevonását a modellbe. $p=1$ esetben nem korrigálunk.



A többváltozós lineáris regresszió

Modell-építési technikák

Egy tipikus többváltozós regressziós problémánál adott az Y célváltozó, és a magyarázó változók közül tudjuk, melyek azok, amelyek befolyásolják a Y értékét. Ha minden lehetséges kombinációt megvizsgáljuk, akkor összesen 15 modell kellene illesztünk!

Már 4 változó esetén 15 modellt kellene illesztünk!

modellillesztést kellene elvégeztünk!

1	1
2	3
3	7
4	15
5	31
6	63
7	127
8	255
9	511
10	1023
11	2047
12	4095
13	8191
14	16383
15	32767
16	65535
17	131071
18	262143
19	524287
20	1048575
21	2097151
22	4194303
23	8388607
24	16777215
25	33554431
26	67108863
27	134217727
28	268435455
29	536870911
30	1073741823
40	1,09951E+12
50	1,12585E+15
100	1,26765E+30

regressziós problémánál adunk $X_1, X_2, \dots, X_p$ magyarázó változókat, és szeretnénk azt sem tudni, melyek azok, amelyek bekerülnek, és melyek nem. Ha mindenképp beadjuk őket a modellbe, akkor ki akarnánk próbálni, hogy melyik a legjobb!



A többváltozós lineáris regresszió

Modell-építési technikák

Nyilván szűkíteniünk kell kell az illesztendő modellek számát!

Alkalmazhatjuk az **ENTER** eljárást, amelyben azokat a magyarázó változókat vesszük be a változólistából a modellbe, amely változókat szeretnénk, hogy benne legyenek. Ezeket a modelleket *utólag* értékelni kell a meghatározottsági együttható nagysága, és a regressziós együtthatók szignifikancia szintje alapján. A módosításokkal újra el kell végezni az illesztést.



A többváltozós lineáris regresszió

Modell-építési technikák

Automatikus modellépítési technikák:

- **STEPWISE**
- **FOREWARD**
- **BACKWARD**
- **REMOVE**

A felhasználónak csak az indulási magyarázó változó listát kell specifikálnia, az SPSS program ebből választva állít elő „jó” modelleket, amik közül választhatunk „végső” megoldást.



A többváltozós lineáris regresszió

A parciális F-próba

Tegyük fel, hogy bevontuk a p -edik magyarázó változót a modellbe. Ha az új változó magyarázó ereje elhanyagolható, akkor az alábbi statisztika 1, $n-p-1$ szabadságfokú Fisher-eloszlást követ:

$$F_p = \frac{R^2 - R_0^2}{1 - R^2} \cdot (n - p - 1)$$

R^2 az új p változós modell meghatározottsági együtthatója,

R_0^2 a régi $p-1$ változós modell meghatározottsági együtthatója,



A többváltozós lineáris regresszió

A parciális F-próba

A p -edik változót akkor vonjuk be a modellbe, ha

$$\frac{K_\varepsilon \cdot (1 - R^2)}{(n - p - 1)} < R^2 - R_0^2$$

ahol

K_ε olyan kritikus érték, hogy: $P(F_{1,n-p-1} < K_\varepsilon) = 1 - \varepsilon$



A többváltozós lineáris regresszió

FOREWARD modell-építés

Alulról építkező modellépítési eljárás. Minden modellépítési lépésben a listából azt a változót vonjuk be, amely F-tesztjéhez a legkisebb ε szint tartozik. A bevonási folyamat addig tart, amíg ez a legkisebb ε szint egy beállított PIN korlát alatt marad. Előnye, hogy viszonylag kevés magyarázó változó lesz a modellben, így könnyebb a modellt értelmezni.



A többváltozós lineáris regresszió

BACKWARD modell-építés

Felülről lebontó eljárás. Kezdetben az összes változót berakjuk a modellbe. Minden lépésben azt a változót hagyjuk el a modellből, amelynél parciális F-próbánál a legnagyobb ε érték tartozik. Akkor állunk meg, ha az előre beállított POUT küszöbérték alá megy ez az ε . A BACKWARD modellépítéssel viszonylag sok magyarázó változó marad benn a modellben.



A többváltozós lineáris regresszió

STEPWISE modell-építés

A FOREWARD eljárást úgy módosítjuk, hogy minden lépésben ellenőrizzük a modellbe korábban már bevont változókhoz tartozó ε szignifikancia-szintet, és azt elhagyjuk, ahol ez a szint nagyobb mint POUT. Nem kerülünk végtelen ciklusba, ha $PIN < POUT$. (Szokásos beállítás: $PIN=0,05$ és $POUT=0,10$.)



A többváltozós lineáris regresszió

REMOVE modell-építés

A REMOVE eljárás az ENTER beállításából indul ki, egyszerre hagy el változókat a modellből, összehasonlításként csak a konstans tagot tartalmazó modell eredményeit közli.



A többváltozós lineáris regresszió

Multikollinearitás

Multikollinearitáson a magyarázó változók között fellépő lineáris kapcsolat meglétét értjük. A multikollinearitás jelenléte rontja a modell értékelhetőségét.

A multikollinearitás mérőszámai:

- tolerancia
- variancia infláló faktor (VIF)
- kondíciós index (CI)
- variancia hányad



A többváltozós lineáris regresszió

A multikollinearitás mérőszámai 1.

tolerancia azt méri, hogy az i -edik magyarázó változót az összes többi milyen szorosan határozza meg. A nullához közeli tolerancia jelenti azt, hogy közel függvényyszerű kapcsolat van a magyarázó változók között. Értéke $1-R_i^2$, ahol R_i az i -edik változónak a többivel vett lineáris regressziójának a korrelációs együtthatója, a többszörös korrelációs együttható.

A **variancia infláló faktor (VIF)** a tolerancia reciproka: $VIF=1/(1-R_i^2)$. Ezért, ha a magyarázó változók között szoros a kapcsolat, VIF végtelen nagy is lehet. Ha a magyarázó változók korrelálatlanok, a VIF értéke 1.



A többváltozós lineáris regresszió

A multikollinearitás mérőszámai 2.

A **kondíciós index (CI)** a magyarázó változók korrelációs mátrixának sajátértékeiből számolt statisztika. A legnagyobb és legkisebb sajátértékek hányadosának négyzetgyöke. A $CI > 15$ esetében megállapítható az erős kollinearitás.

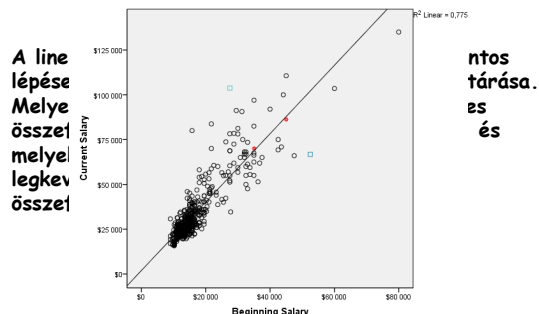
$$CI = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

Variancia hányad is utalhat multikollinearitásra. Ha egy-egy nagy kondíciós index sorában több regressziós együtthatónak van magas variancia hányada. A regressziós együtthatók varianciáit a sajátértékek között szétosztjuk.



A többváltozós lineáris regresszió

A becslést befolyásoló pontok feltárása



A többváltozós lineáris regresszió

A becslést befolyásoló pontok feltárása

A $\underline{Y}$ célváltozó és a lineáris becslés közötti kapcsolat:

$$\tilde{Y} = \underline{X} \cdot \underline{B} = \underline{X} \cdot (\underline{X}^T \underline{X})^{-1} \underline{X}^T \cdot \underline{Y} = \underline{H} \cdot \underline{Y} \quad \underline{H} = \underline{X} \cdot (\underline{X}^T \underline{X})^{-1} \underline{X}^T$$

A becslés hibavektora, maradékösszeg, regressziós összeg:

$$\underline{e} = \underline{Y} - \tilde{Y} = (\underline{E} - \underline{H}) \underline{Y} \quad SSE = \underline{Y}^T \cdot (\underline{E} - \underline{H}) \underline{Y} \quad SSR = \underline{Y}^T \underline{H} \cdot \underline{Y} - n \cdot (\bar{y}_n)^2$$



A többváltozós lineáris regresszió

A becslést befolyásoló pontok feltárása

$$\underline{H} = \underline{X} \cdot (\underline{X}^T \underline{X})^{-1} \underline{X}^T \quad \text{a leverage (hatalom) vagy hat mátrix}$$

A mátrix szimmetrikus, h_{ii} diagonális elemei azt mutatják, hogy az i -edik eset mekkora hatást fejt ki a regressziós becslésre.

$$h_{ii} = \underline{x}_i^T (\underline{X}^T \underline{X})^{-1} \underline{x}_i, \quad \text{ahol } \underline{x}_i \text{ az } i\text{-edik esetvektor}$$

$$\sum_{i=1}^n h_{ii} = p + 1 \quad \frac{1}{n} \leq h_{ii} \leq 1$$



A többváltozós lineáris regresszió

A becslést befolyásoló pontok feltárása

Az i -edik eset befolyása átlagos, $h_{ii} < \frac{p+1}{n}$ ezek a tipikus esetek!

Az i -edik eset befolyása jelentős, ha $h_{ii} > 2 \cdot \frac{p+1}{n}$

Ha $h_{ii} - \frac{1}{n} < 0,2$ az i -edik eset bevonható az elemzésbe

Ha $0,2 \leq h_{ii} - \frac{1}{n} < 0,5$ kockázatos az i -edik eset bevonása

$h_{ii} - \frac{1}{n} \geq 0,5$ az i -edik esetet ki kell hagyni, „outlier” pont



A többváltozós lineáris regresszió

A maradéktagok (reziduálisok) elemzése

$$\text{Közönséges reziduális: } e_i = Y_i - \hat{Y}_i$$

$$\text{Törölt reziduális: } e_{(i)i} = Y_i - \hat{Y}_{(i)i} = \frac{e_i}{1 - h_{ii}}$$

$$\text{Standardizált reziduális: } z_i = \frac{e_i}{\sqrt{\sum_{i=1}^n e_i^2}} \sqrt{n - p - 1}$$

$$\text{Belsőleg studentizált reziduális: } r_i = \frac{e_i}{\sqrt{\sum_{i=1}^n e_i^2}} \frac{\sqrt{n - p - 1}}{\sqrt{1 - h_{ii}}}$$

A lineáris becslés elkészítésekor nem számolunk az i -edik esettel, „töröljük”.

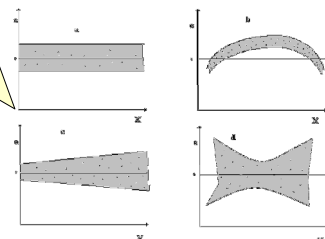


A többváltozós lineáris regresszió

A maradéktagok (reziduálisok) elemzése

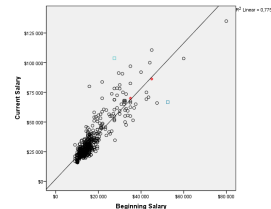
Heteroszkedaszticitás:

A maradéktagok nulla szint körüli szóródásának lehetséges típusai
a.) a szóródás megfelel a lineáris modellnek,
b.) nem a lineáris modellhez tartoznak a maradéktagok,
c.) a szóródások nem azonosak,
d.) a hibátagok nem függetlenek egymástól.



Példa kétváltozós lineáris regresszióra

Keressünk lineáris összefüggést az *employee data* állományban a kezdőfizetés és a jelenlegi fizetés között!



Példa kétváltozós lineáris regresszióra



Példa kétváltozós lineáris regresszióra

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.880 ^a	.775	.774	\$8,115.356

a. Predictors: (Constant), salbegin Beginning Salary

ANOVA^a

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	1,066E11	1	1,066E11	1622,118	,000 ^a
Residual	3,109E10	472	6,588E7		
Total	1,379E11	473			

a. Predictors: (Constant), salbegin Beginning Salary

b. Dependent Variable: salary Current Salary

Coefficients^a

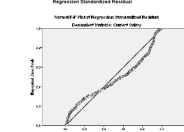
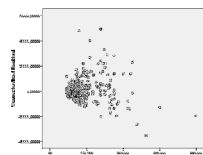
Model	Unstandardized Coefficients		Standardized Coefficients		Sig.
	B	Std. Error	Beta	t	
1 (Constant)	1926,206	888,680		2,170	,031
salbegin Beginning Salary	1,909	,047	,880	40,276	,000

a. Dependent Variable: salary Current Salary



Példa kétváltozós lineáris regresszióra

a maradéktagok



Heteroszkedaszticitás jelensége megfigyelhető: nagyobb X-hez nagyobb szórás tartozik!



Példa kétparaméteres nemlineáris regresszióra

Keressünk nemlineáris kapcsolatot Cars állományban a lóerő és a fogyasztás között!

