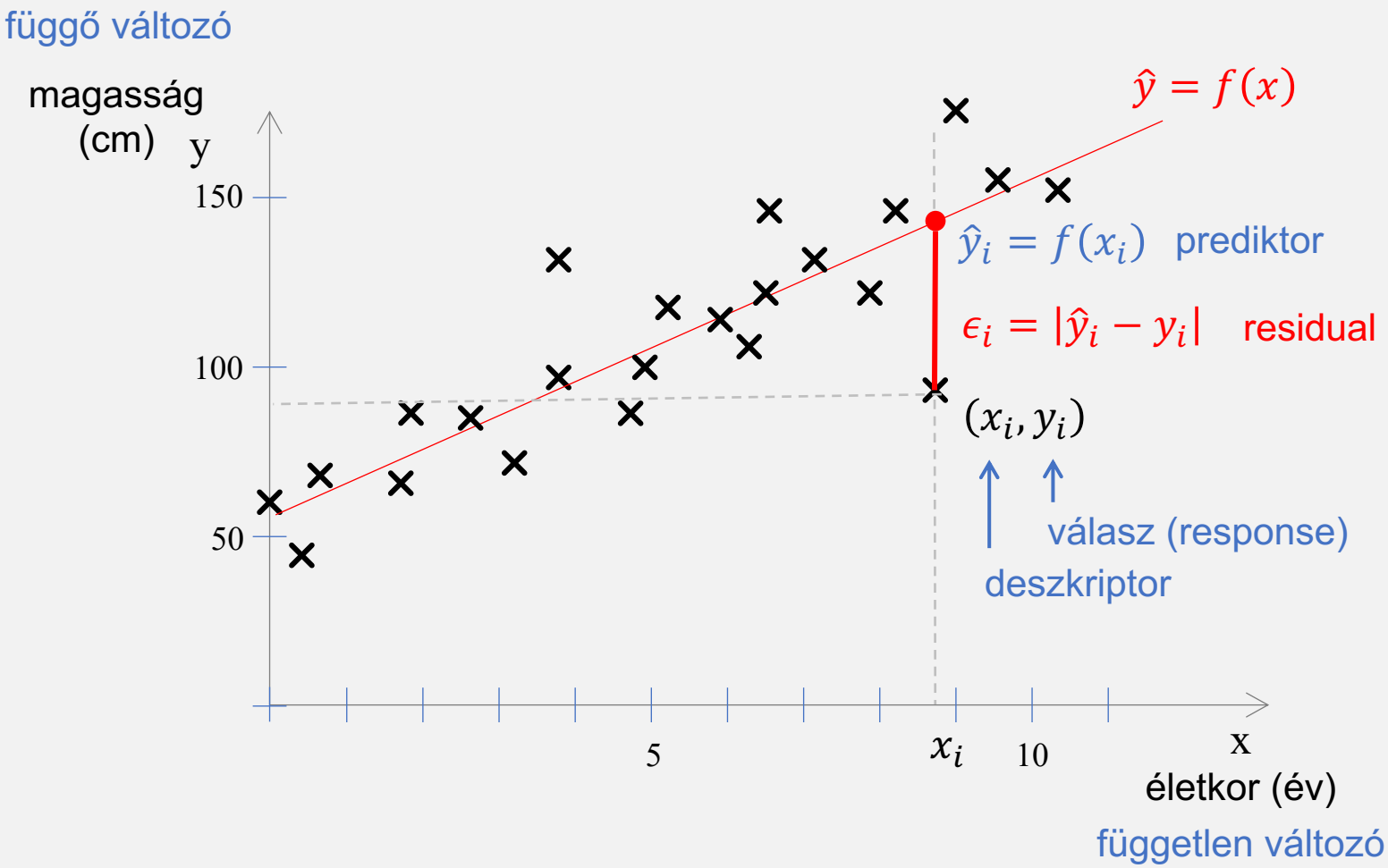
REGRESSZIÓ, PREDIKCIÓ

REGRESSZIÓ

- Regresszió
  - A regresszió egy statisztikai technika két, vagy több változó közötti összefüggés megállapítására
  - Milyen kapcsolat áll fenn két, vagy több változó között?
- Adatsor
  - $(x_{1i}, x_{2i}, \dots x_{pi}) \rightarrow y_i$
  - Függelen változó:$(x_{1i}, x_{2i}, \dots x_{pi})$
  - Függő változó: y_i
- Feladat
  - Modellalkotás
    - Akár több model is létezhet
    - Model kiértékelés
    - Alkalmazandó model választás

SZEMLÉLTETÉS



REGRESSZIÓ TÍPUSAI

- A független változók száma szerint
  - Egyszerű regresszió (Simple Regression)
  - Többváltozós regresszió (Multivariate regression)
- Az összefüggés szerint
  - Lineáris regresszió (Linear Regression)
  - Nemlineáris regresszió
    - Exponenciális (Exponential)
    - Hatványfügvénnyel leírható
    - Hatványpolinommal lerható (Polinomial)
    - Hiperbolikus (Hyperbolic)
    - … Tetszőleges függvénnyel adott
    - Logisztikus (Logistic)

GÉPI TANULÁS ÉS REGRESSZIÓ

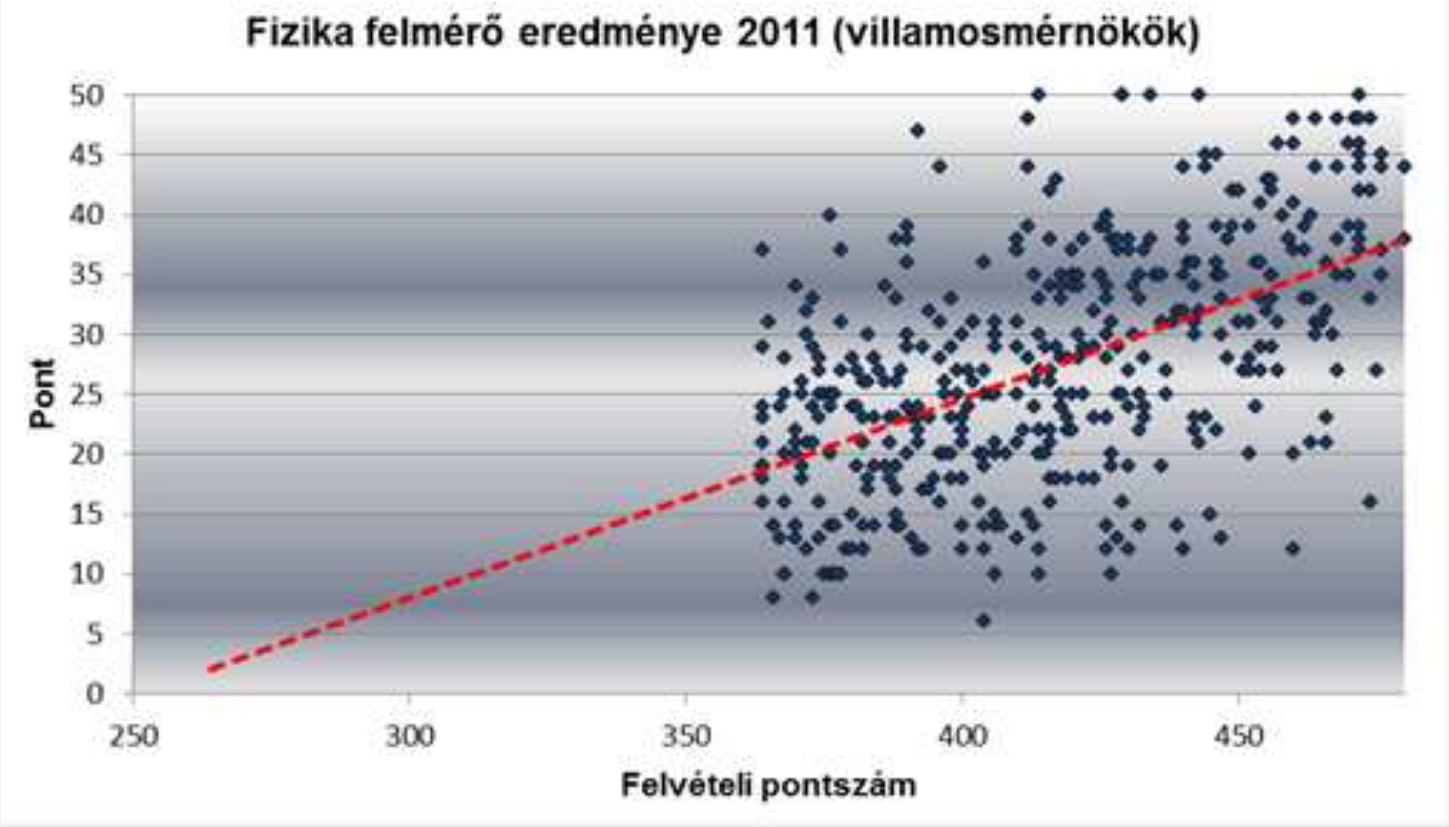
- Teljes adathalmaz felbontása
  - Tanító adathalmaz
    - Modellek megfogalmazása
  - Validáló adathalmaz
    - Model kiválasztása
    - Paraméterek meghatározása
  - Teszt adathalmaz
    - Kiválasztott modell kiértékelése

$$Adathalmaz = Tanító\ (70\%) + Validáló(15\%) + Teszt(15\%)$$

LINEÁRIS REGRESSZIÓ

- Egyszerű lineáris regresszió
  - Lineáris kapcsolat meghatározása y és x között
    - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - x_i értékek felelősek $\hat{y}_i$ értékek prediktálásáért
    - $\hat{y}_i = \beta_0 + \beta_1 x_i$
  - A független változót körültekintően kell megválasztani
- Többszörös lineáris regresszió
  - y több x változó lineáris kombinációja
    - $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \epsilon_i$
  - Predikció
    - $\hat{y}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i}$
- Az együtthatók (koefficiensek) meghatározása
  - $\sum_{i=1..N} \epsilon_i^2 = \sum (y_i - \hat{y}_i)^2$ minimalzálása
  - Least Squares (LS) eljárások

ELHAMARKODOTT ALKALMAZÁS



MODELL MEGFELELŐSÉG

- Feltételek
  - Összefüggés: független és függő kapcsolat között kell létezzen összefüggés
  - Függetlenség: Az egyes értékekre számolt residual-ok értéke független legyen egymástól
  - Homogenitás: Az teljes értelmezési tartományon a residual-ok szórása homogén legyen
  - Normalitás: A gyűjtött minták hibái normális eloszlást kövessenek
- Változók közötti összefüggés
  - → Összefüggés
- Hibatag eloszlása
  - → Normalitás
- Modell kiértékelés **`plot()`** függvénnyel
  - → Homogenitás

MODELL MEGFELELŐSÉG

- Mérőszámok
  - MAE (Mean Absolute Error)
    - $MAE = \frac{1}{n}\sum_{i=1..n}|y_i - \hat{y}_i|$
  - MAPE (Mean Absolute Percentage Error)
    - $MAPE = \frac{1}{n}\sum_{i=1..n}|\frac{y_i - \hat{y}_i}{y_i}|$
  - RMSE (Route Mean Square Error)
    - $RMSE = \sqrt{\frac{1}{n}\sum_{i=1..n}(y_i - \hat{y}_i)^2}$
    - Az RMSE a mintákra számolt szórás értéke
  - R négyzet
    - $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$
    - A minták és prediktorok közötti koreláció négyzete

PREDIKCIÓ

- A predikció a regressziós modell alapján való $\hat{y}_j$ érték becslése a minták között nem szereplő x_j értékhez
  - $\hat{y}_j = f_{model}(x_j)$
  - **`predict()`**
- Intervallum becslés
  - Becslési (**`prediction`**) intervallum
    - Bizonytalanság egy érték körül
    - Adott x_j értékek 95%-ához tartozó y_j érték benne lesz az intervallumban
    - Általában erre vagyunk kíváncsiak
  - Konfidencia (**`confidence`**) intervallum
    - Bizonytalanság a predikált átlag körül
    - Adott x_i értékhez tartozó y_j értékek átlaga 95% valószínűséggel az intervallumba fog esni

NEMLINEÁRIS REGRESSZIÓ

- Feladat
  - Az R nemlineáris regresszióanalízise a nemlineáris függvény felépítésének folyamata.
- Jellemző összefüggések
  - Exponenciális regresszió
    - $y_i = ab^{x_i} + \epsilon_i$
  - Polinomiális regresszió
    - $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$
  - Logisztikus regresszió
    - Gyakorlatias folytonos értékek predikálása
    - Binomiális értékek predikálása
    - Kategorikus értékek predikálása
    - $y_i = A \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$

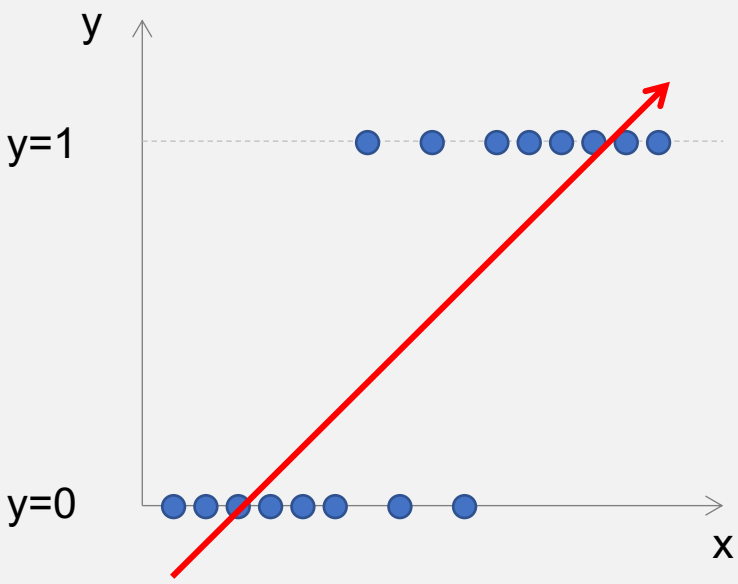
NEMLINEÁRIS REGRESSZIÓ

- Visszavezetés lineáris regresszióra

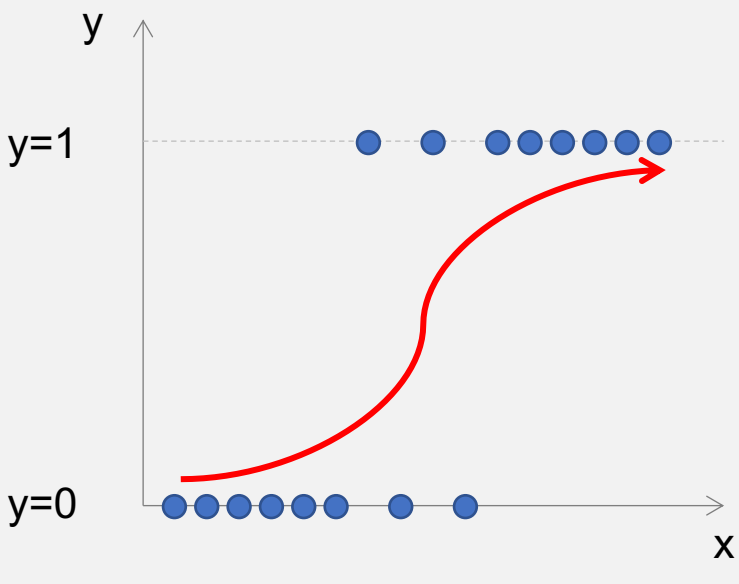
$$\hat{y} = ab^x$$
$$\log y = \log a + \mathrm{x} \cdot \log b$$
$$y' = a' + b' \cdot x'$$

  - Változók
    - $y' = \log y, \ldots x' = x$
  - Együtthatók
    - $a' = \log a, \ldots b' = \log b$
- Általános megoldás
  - **`nlm(<formula>, start = list(), data = <data>)`**

LOGISZTIKUS REGRESSZIÓ



Lineáris regresszió



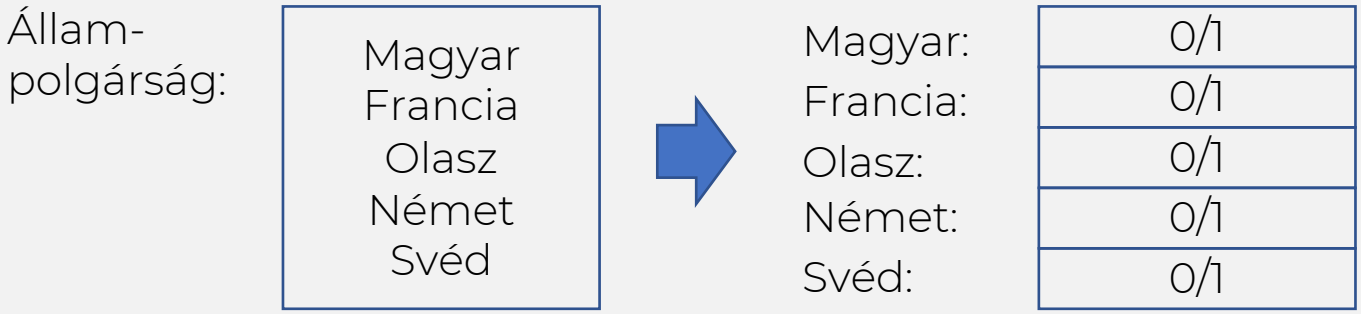
Logisztikus regresszió

BINOMIÁLIS ÉRTÉKEK

- Döntési valószínűség
  - $p(x) = P(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$
- Esélyességi ráta (odds ratio)
  - $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$
- Logit függvény
  - $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$
- Megvalósítás
  - **`glm(…, family = "binomial")`**

KATEGORIKUS VÁLTOZÓK

- Kategorikus változók felbontása bináris változókra
  - One-hot codig
    - Minden kategóriához egy-egy bináris változót
    - Bináris változó az adott kategóriába való tartozást jelöli



  - Dummy codig
    - Egyel kevesebb bináris változó
    - Az összes 0 értékek egy kijelölt kategóriát jelöl