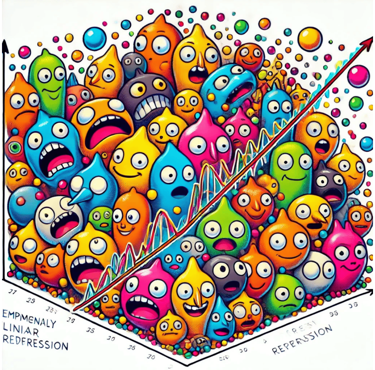
Regresszióanalízis: egyváltozós lineáris regresszió

Matematikai Statisztika
2024. október 21.



Egyszerű lineáris regresszió - Adatok és modellfeltevések

Függvénycsalád megadása

Tegyük fel, hogy rendelkezésünkre áll egy minta az $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ adatokkal, ahol:

- Az X_i és Y_i értékek az ismeretlen eloszlású X és Y valószínűségi változókból származnak.
- Szeretnénk, hogy egy lineáris összefüggést találjunk közöttük a következő formában:

$$Y_i = a + bX_i + \varepsilon_i,$$

  ahol a és b az ismeretlen paraméterek, ε_i pedig a hibatag.

Mi a célunk?

A regressziós modell célja, hogy **minimalizáljuk** az eltéréseket a becsült Y és a tényleges Y értékek között, azaz a Y_i és a $a + bX_i$ közötti különbséget. A legjobb lineáris illesztés megtalálásához ezt a különbséget négyzetre emeljük és minimalizáljuk:

$$\sum_{i=1}^{n} (Y_i - (a + bX_i))^2.$$

Lineáris modell - Hibatagok feltételei

A hibatagokra vonatkozó feltételek

A lineáris regresszió használatához a következő feltételezéseket tesszük a **hibatagokra** (ε_i):

- **Nulla várható érték**: A hibatagok várható értéke 0, azaz nincsenek szisztematikus eltérések:

$$\mathbb{E}(\varepsilon_i) = 0.$$

- **Homoszkedaszticitás**: A hibatagok varianciája állandó minden megfigyelésre, azaz:

$$\mathrm{Var}(\varepsilon_i) = \sigma^2.$$

  Ezt nevezzük **homoszkedaszticitásnak**.

- **Függetlenség**: A hibatagok egymástól függetlenek:

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j.$$

- **Függetlenség a magyarázó változótól**: Az X_i és ε_i függetlenek egymástól:

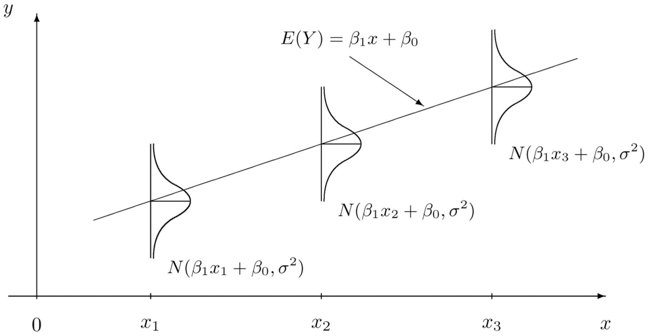
$$\mathrm{Cov}(X_i, \varepsilon_i) = 0.$$

- **Normális eloszlás**: A hibatagok normális eloszlásúak:

$$\varepsilon_i \sim N(0, \sigma^2).$$

Bár vannak modellek, amelyek kevesebb feltételt követelnek meg, mi itt az összes klasszikus feltételt alkalmazzuk, hogy a lehető legjobb eredményt kapjuk.

Lineáris modell



Normálegyenletek és a becslések levezetése

A legkisebb négyzetek módszere

A legkisebb négyzetek módszerével a célunk, hogy minimalizáljuk a négyzetes eltéréseket a tényleges Y_i és a becsült $a + bX_i$ értékek között. A veszteségfüggvény:

$$V(a, b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2.$$

Ezt kell deriválni a és b szerint, hogy megtaláljuk az optimális értékeket.

Normálegyenletek és a becslések levezetése

Normálegyenletek levezetése

A parciális deriváltak segítségével kapjuk az úgynevezett **normálegyenleteket**, amelyeket a következő lépésekben vezetünk le:

- Az a szerinti deriválás:

$$\frac{\partial V(a,b)}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n}(Y_i - a - bX_i)^2 = -2\sum_{i=1}^{n}(Y_i - a - bX_i) = 0,$$

  amit átrendezve kapjuk:

$$\sum_{i=1}^{n}(Y_i - a - bX_i) = 0.$$

- A b szerinti deriválás:

$$\frac{\partial V(a,b)}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{n}(Y_i - a - bX_i)^2 = -2\sum_{i=1}^{n}X_i(Y_i - a - bX_i) = 0,$$

  amit átrendezve kapjuk:

$$\sum_{i=1}^{n}X_i(Y_i - a - bX_i) = 0.$$

Ezek az egyenletek alkotják a **normálegyenleteket**, amelyek segítségével kiszámíthatjuk a és b értékeit.

Az együtthatók becslése

Normálegyenletek megoldása

A normálegyenletek megoldásával a következő képleteket kapjuk az a és b paraméterekre:

$$na + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

és

$$a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i.$$

Ezek az egyenletek lineáris egyenletrendszert alkotnak a és b ismeretlenekkel. A megoldásokat átrendezve és egyszerűsítve kapjuk az alábbi becsléseket.

Normálegyenletek és a becslések levezetése

Az a és b paraméterek becslése

A megoldásokat a következőképpen kapjuk meg:

$$\hat{b} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

Más módszerek

Ezek a **legkisebb négyzetek módszerével** kapott becslések, amelyek minimalizálják a négyzetes eltéréseket. Ezen becsléseken kívül más módszerekkel, például maximum likelihood vagy robusztus módszerekkel is meghatározhatjuk az együtthatókat.

Konstans (intercept)

A konstans azt mutatja meg, hogy mekkora lenne a függő változó (Y) értéke, ha az összes független változó (X) értéke nulla lenne. Például, ha a regresszió egy ház árát próbálja becsülni, akkor a konstans mutatja meg az alapárat.

Meredekség (slope)

Az X változóhoz tartozó együttható. Ez az érték azt mutatja meg, hogy ha X értéke egységnyivel nő, akkor hogyan változik a Y értéke. A meredekség az X és Y közötti kapcsolat erősségét és irányát mutatja.

Példa

Egy regresszióban, ahol Y a ház árát, X pedig a ház méretét jelöli, egy $\hat{b} = 200$ azt jelenti, hogy minden extra négyzetméterrel 200 egységgel nő az ár, azaz a négyzetméterár.

Lineáris regresszió - Elaszticitás

Elaszticitás definíciója

Az elaszticitás egy aránymutató, amely a magyarázó változó és az eredményváltozó relatív változását méri. Az elaszticitás azt mutatja meg, hogy az X változó 1%-os növekedése mekkora százalékos változást eredményez Y-ban.

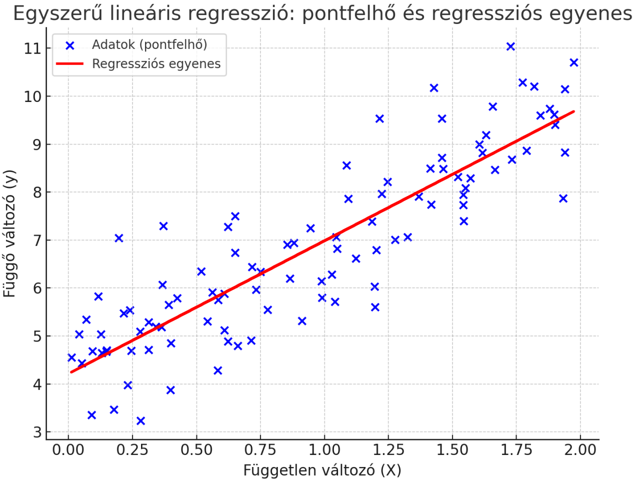
Elaszticitás:

$$E\ell(\hat{Y}|X) = \frac{\hat{b}X}{\hat{a} + \hat{b}X}.$$

Interpretáció

Ha az elaszticitás értéke 0.5, akkor az X változó 1%-os növekedése Y értékének 0.5%-os növekedésével jár. Az elaszticitás segít a regressziós modell relatív változásainak mérésében, ami különösen hasznos, ha a változók különböző mértékegységekben vannak.

Lineáris regressziós egyenes



Lineáris regresszió - Előrejelzés és rajz

Előrejelzés lineáris regresszióval

A lineáris regresszió segítségével az X változó új értékei alapján meg tudjuk becsülni a függő változó Y várható értékét. Az előrejelzés lépései a következők:

- Adott egy új X^* érték, azaz egy új független változó megfigyelés,
- A becsült lineáris modell alapján az Y^* függő változóra a következőképpen adhatunk előrejelzést:

$$\hat{Y}^* = \hat{a} + \hat{b}X^*,$$

  ahol:
  - $\hat{a}$: az Y-tengelymetszet becsült értéke (intercept),
  - $\hat{b}$: a meredekség becsült értéke,
  - X^*: az új független változó megfigyelt értéke.

Ez az egyenlet lehetővé teszi, hogy az új X^* érték alapján megjósoljuk a hozzá tartozó Y^* függő változót.

Lineáris regresszió - Rajz

Rajz

Hiba varianciájának becslése a lineáris regresszióban

Miért fontos a hiba varianciája?

A hiba varianciája (σ^2) kifejezi, hogy milyen mértékben térnek el a tényleges Y_i értékek a regressziós egyenes által becsült $a + bX_i$ értékektől. Ez a variancia megmutatja, hogy a modell milyen jól illeszkedik az adatokra, illetve hogy mennyire pontosak a becslések.

A hiba varianciájának becslése

A hiba varianciáját a következőképpen becsüljük meg:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2,$$

ahol:

- n a megfigyelések száma,
- $\hat{a}$ és $\hat{b}$ a becsült együtthatók,
- $Y_i - \hat{a} - \hat{b}X_i$ az i-edik megfigyelés reziduálja, azaz a tényleges Y_i és a becsült Y érték közötti eltérés.

Az $n - 2$ szerepe

Az $n - 2$ az úgynevezett **szabadságfok**, amely a két paraméter ($\hat{a}$ és $\hat{b}$) becslésére történő illesztést korrigálja. Ez biztosítja, hogy a becslés ne legyen torzított, és figyelembe vegye a modellezés szabadságfokait.

Az együtthatók becslésének eloszlása

Feltételek és cél

A lineáris regressziós modell során feltételezzük, hogy a hibatagok (ε_i) függetlenek, azonos eloszlásúak, és normális eloszlásúak, azaz:

$$\varepsilon_i \sim N(0, \sigma^2).$$

Ezek a feltételek biztosítják, hogy a regressziós együtthatók becslései is normális eloszlásúak legyenek. Ezen becslések segítségével meghatározhatjuk a paraméterek pontosságát és bizonytalanságát.

Az együtthatók becslésének eloszlása

Az együtthatók becsléseinek eloszlása

- **Meredekség becslése** ($\hat{b}$): A $\hat{b}$ becslés normális eloszlású a következő paraméterekkel:

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{\sum(X_i - \overline{X})^2}\right),$$

  ahol σ^2 a hiba varianciája, és $\overline{X}$ a független változó átlaga.
- **Y tengelymetszet becslése** ($\hat{a}$): Hasonlóképpen, $\hat{a}$ becslése is normális eloszlású:

$$\hat{a} \sim N\left(a, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right)\right),$$

  ahol n a minta mérete, σ^2 a hiba varianciája.
- Mindkét becslés **torzítatlan**, azaz az eloszlásaik várható értéke a valódi paraméterekkel (a és b) egyezik meg.

Összegzés és alkalmazás

A normális hibatagok feltételezése alapján az a és b becslései normális eloszlásúak. Ezek az eloszlások segítenek a **konfidencia intervallumok** és a **statisztikai tesztek** meghatározásában, amelyekkel megadhatjuk a becslések bizonytalanságát és pontosságát.

A meredekség becslésének standard hibája ($\hat{b}$)

A meredekség becslésének ($\hat{b}$) standard hibája:

$$\mathrm{SE}(\hat{b}) = \sqrt{\frac{\sigma^2}{\sum(X_i - \overline{X})^2}}$$

ahol:

- σ^2: a hibák varianciája,
- $\sum(X_i - \overline{X})^2$: a független változó szórásnégyzete.

Ez a kifejezés azt mutatja meg, hogy milyen mértékben változhat a becsült meredekség ($\hat{b}$) értéke a mintavételezés következtében.

Lineáris regresszió - $\hat{a}$ standard hibája

A tengelymetszet becslésének standard hibája ($\hat{a}$)

A tengelymetszet ($\hat{a}$) standard hibája:

$$\mathrm{SE}(\hat{a}) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right)}$$

ahol:

- n: a minta mérete,
- $\overline{X}$: a független változók átlaga.

Ez a kifejezés azt mutatja, hogy mennyire bizonytalan a tengelymetszet becslése a mintavétel következtében.

Intervallumbecslés az együtthatókra

A regressziós modellben az a és b paraméterek becslésén túl fontos **intervallumbecslést** is adnunk, hogy kifejezzük a becslések bizonytalanságát. Az intervallum azt mutatja meg, hogy adott megbízhatósági szinten (például 95%) az együtthatók valódi értékei milyen tartományban helyezkedhetnek el.

Intervallumbecslés kiszámítása

A $\hat{a}$ és $\hat{b}$ becsléseinek **intervallumbecslését** a következőképpen adjuk meg:

- Az a becslésére adott $1-\varepsilon$ megbízhatósági szinten:

$$\hat{a} \pm t_{n-2,1-\varepsilon/2} \cdot SE(\hat{a}),$$

  ahol $SE(\hat{a})$ az a becslésének standard hibája, és $t_{n-2,1-\varepsilon/2}$ a Student-féle t-eloszlás kvantilis értéke $n-2$ szabadságfokkal.

- A b becslésére hasonlóképpen:

$$\hat{b} \pm t_{n-2,1-\varepsilon/2} \cdot SE(\hat{b}),$$

  ahol $SE(\hat{b})$ a b becslésének standard hibája.

Lineáris regresszió - Intervallumbecslés az együtthatókra

Hogyan értelmezzük az intervallumbecslést?

Az intervallumbecslés azt mutatja meg, hogy $1 - \varepsilon$ megbízhatósági szinten (például 95%-on) milyen tartományban várható az együtthatók valódi értéke. Ez a tartomány kifejezi, hogy mennyire bizonytalan az együtthatók becslése. Minél kisebb a standard hiba (*SE*), annál szűkebb az intervallum, és annál pontosabb a becslés.

Reggressziós egyenes - Konfidencia intervallum

Konfidencia intervallum a regressziós egyenesre

A regressziós egyenes becsült értékei köré konfidencia intervallumot számítunk, hogy megbecsüljük, milyen tartományban várható az Y érték.
Konfidencia intervallum formulája egy adott X pontnál:

$$\hat{Y} \pm t_{\alpha/2} \cdot \mathrm{SE}(\hat{Y})$$

ahol:

- $\hat{Y} = \hat{a} + \hat{b}X$: az X értéknél becsült Y,
- $t_{\alpha/2}$: a t-eloszlás kritikus értéke adott szignifikanciaszinten,
- $\mathrm{SE}(\hat{Y})$: a becsült értékek standard hibája, amely tartalmazza a hibát a modellből.

Mit mutat a konfidencia intervallum?

A konfidencia intervallum megmutatja, hogy az adott X értéknél milyen tartományban várható az Y érték a választott szignifikancia szinten (pl. 95%-os konfidenciaszinten).

Előrejelzési intervallum kiszámítása

Előrejelzési intervallum definíciója

Az előrejelzési intervallum azt mutatja meg, hogy az új megfigyeléshez ($Y_{\text{űj}}$) tartozó érték milyen tartományba esik, ha ismerjük az új $X_{\text{új}}$ értéket. Az intervallum nemcsak a modell bizonytalanságát, hanem az új megfigyelés varianciáját is figyelembe veszi.

Az előrejelzési intervallum kiszámítása

Az $X_{\text{űj}}$-hez tartozó $Y_{\text{űj}}$ előrejelzés intervalluma $1-\varepsilon$ megbízhatósági szinten a következőképpen adható meg:

$$Y_{\text{új}} \pm t_{n-2,1-\varepsilon/2} \cdot \sqrt{SE_{\text{ill}}^2 + \hat{\sigma}^2}$$

ahol:

- $t_{n-2,1-\varepsilon/2}$ a Student-féle t-eloszlás kvantilise $n-2$ szabadságfokkal,
- SE_{ill} az illesztés standard hibája, amelyet a becsült egyenes bizonytalansága ad:

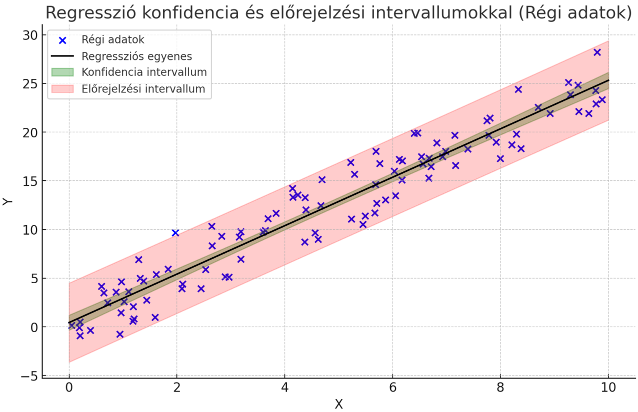
$$SE_{\text{ill}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_{\text{űj}} - \overline{X})^2}{\sum (X_i - \overline{X})^2}},$$

- $\hat{\sigma}^2$ a hibatagok becsült varianciája, amely a modell hibáját fejezi ki.

Előrejelzési intervallum kiszámítása

Összegzés

Az előrejelzési intervallum szélesebb, mint a konfidenciaintervallum, mert figyelembe veszi az új megfigyelés körüli bizonytalanságot, valamint az együtthatók becslési bizonytalanságát. Ezzel megmutatja, hogy az adott $X_{új}$ értékhez milyen tartományban várhatjuk $Y_{új}$-t.



23/39

Konfidencia és előrejelzési intervallum összehasonlítása

Konfidencia intervallum

- A regressziós egyenes becsült átlagos értékeinek tartományát adja meg egy adott X pontban.
- Csak a regressziós modell paramétereinek bizonytalanságát veszi figyelembe.
- Szűkebb, mivel kevesebb bizonytalanságot tartalmaz.
- Alkalmazható a modell paramétereinek pontosságának értékelésére.

Előrejelzési intervallum

- Az új megfigyelések Y értékeinek várható tartományát adja meg egy adott X pontban.
- Figyelembe veszi a modell bizonytalanságát és az új megfigyelések természetes variabilitását.
- Szélesebb, mivel több bizonytalansági tényezőt tartalmaz.
- Alkalmazható új, jövőbeli adatok Y értékeinek becslésére.

Összegzés

A konfidencia intervallum a modell pontosságát mutatja, míg az előrejelzési intervallum a jövőbeli megfigyelések várható variabilitását becsüli meg.

A lineáris regresszió sorrendje

1. Az együtthatók becslése:

- A regressziós egyenlet együtthatóit (a és b) a **legkisebb négyzetek módszerével** becsüljük.
- Ezzel pontbecslést kapunk a paraméterekre, és megkapjuk a legjobban illeszkedő egyenest.

2. A hiba becslése:

- Az együtthatók alapján kiszámítjuk a **hibatagok varianciáját** ($\hat{\sigma}^2$).
- Ez megmutatja, mennyire térnek el a megfigyelt Y értékek a becsült regressziós egyenestől.

3. Az együtthatók eloszlásának becslése:

- Az együtthatók becsléseinek szórását ($SE(\hat{a})$ és $SE(\hat{b})$) is meghatározzuk.
- Ezzel számszerűsíthetjük a becslések körüli bizonytalanságot.

4. Konfidenciaintervallum az együtthatókra és a regressziós egyenesre:

- A becslések alapján **konfidenciaintervallumokat** adhatunk meg az együtthatókra (pl. 95%-os megbízhatósággal).
- A **konfidenciaintervallum a regressziós egyenesre** megmutatja, hogy a becsült átlagos értékek milyen bizonytalansággal terheltek.

5. Pont- és intervallumbecslés az előrejelzésre:

- A modell alapján becslést adhatunk egy új $X_{\text{új}}$ értékhez tartozó $Y_{\text{új}}$ értékre.
- Az **előrejelzési intervallum** figyelembe veszi mind az együtthatók becslési bizonytalanságát, mind a hibatagok varianciáját.

Összegzés:

- Először az együtthatók becslése, majd a hibák és eloszlások becslése következik.
- Majd konfidenciaintervallumokat és előrejelzési intervallumokat határozunk meg.

Feladat

Kismacskák magasságát szeretnénk előrejelezni, miután felfedeztük, hogy a kismacskák által elfogyasztott napi tejmennyiség (X) és a magasságuk (Y) között lineáris kapcsolat van. A kismacskák különböző napokon eltérő mennyiségű tejet fogyasztanak, és ennek hatására változik a magasságuk is.
A rendelkezésre álló adatok alapján a lineáris regresszió modelljét alkalmazzuk a következő formában:

$$Y = a + bX + \varepsilon,$$

ahol:

- Y: a kismacskák magassága (cm-ben),
- X: az elfogyasztott tej mennyisége (ml-ben),
- a: az Y-tengelymetszet, azaz a magasság akkor, ha $X = 0$,
- b: a meredekség, azaz mennyivel nő a kismacska magassága minden egyes ml tej után.

Most egy új kismacska van a birtokunkban, aki 120 ml tejet ivott, és szeretnénk megbecsülni a magasságát a meglévő adatok alapján. Számoljuk ki a következőket:

- A regressziós modell paramétereit ($\hat{a}$ és $\hat{b}$),
- A kismacska becsült magasságát ($\hat{Y}^*$),
- A konfidenciaintervallumot a becsült magasságra,
- Az előrejelzési intervallumot az új megfigyelésre.

Kismacskák és a lineáris regresszió - Adatok

1. A kismacskák adatai

A kismacskák eddigi adatai a tejfogyasztásról (X_i) és a magasságról (Y_i):

Napi tejfogyasztás (ml) (X_i)	Magasság (cm) (Y_i)
100	25
150	28
200	31
250	35
300	38

A regressziós egyenlet formája: $\hat{Y} = \hat{a} + \hat{b}X$, ahol $\hat{a}$ az Y-tengelymetszet, $\hat{b}$ pedig a meredekség.

A regressziós egyenes paramétereinek kiszámítása

2. A meredekség ($\hat{b}$) kiszámítása

A meredekség ($\hat{b}$) képlete:

$$\hat{b} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$$

A tejfogyasztás átlaga $\overline{X} = 200\,\mathrm{ml}$, a magasság átlaga $\overline{Y} = 31.4\,\mathrm{cm}$.
A számítások után:

$$\hat{b} = 0.066$$

3. A tengelymetszet ($\hat{a}$) kiszámítása

A tengelymetszet ($\hat{a}$) kiszámítása:

$$\hat{a} = \overline{Y} - \hat{b} \cdot \overline{X}$$

Számítások után:

$$\hat{a} = 18.2$$

Így a regressziós egyenlet:

$$\hat{Y} = 18.2 + 0.066X$$

Az új kismacska magasságának becslése

4. Az új kismacska becsült magassága

Az új kismacska 120 ml tejet ivott, így a becsült magassága:

$$\hat{Y}^* = \hat{a} + \hat{b}X^*$$

ahol $X^* = 120$, így:

$$\hat{Y}^* = 18.2 + 0.066 \times 120 = 26.12\,\text{cm}$$

Konfidenciaintervallum a becsült magasságra

5. Konfidenciaintervallum a becsült magasságra (95%-os szinten)

A becsült magasságra adott konfidenciaintervallum:

$$\hat{Y}^* \pm t_{n-2,1-\varepsilon/2} \cdot SE_{\text{fit}}$$

ahol:

- $t_{3,0.975} = 2.306$,
- $SE_{\text{fit}} = 0.214\,\text{cm}$.

Konfidenciaintervallum:

$$26.12 \pm 2.306 \times 0.214 = [25.63, 26.61]\,\text{cm}$$

Előrejelzési intervallum az új megfigyelésre

6. Előrejelzési intervallum az új megfigyelésre

Az előrejelzési intervallum:

$$\hat{Y}^* \pm t_{n-2,1-\varepsilon/2} \cdot \sqrt{SE_{\text{fit}}^2 + \hat{\sigma}^2}$$

ahol $\hat{\sigma}^2 = 0.444$, a hibatagok becsült varianciája.
Számítások után:

$$26.12 \pm 2.306 \times \sqrt{0.214^2 + 0.444} = [25.24, 27.00]\,\text{cm}$$

Az együtthatók tesztelése

1. A meredekség ($\hat{b}$) tesztelése

Teszteljük a hipotézist, hogy $H_0 : b = 0$ (nincs lineáris kapcsolat) és $H_1 : b \neq 0$.
A próbastatisztika képlete:

$$t = \frac{\hat{b}}{SE(\hat{b})}$$

ahol $SE(\hat{b})$ a meredekség standard hibája, amelyet az alábbi képlettel számítunk ki:

$$SE(\hat{b}) = \frac{\sigma}{\sqrt{\sum(X_i - \overline{X})^2}} = 0.0030$$

Számítás után a próbastatisztika:

$$t = \frac{0.0660}{0.0030} = 22.11$$

A kritikus érték: $t_{3,0.975} = 3.182$. Mivel $t = 22.11 > 3.182$, elutasítjuk H_0-t, és kijelenthetjük, hogy van szignifikáns lineáris kapcsolat a tejfogyasztás és a magasság között.

Az együtthatók tesztelése (folytatás)

2. A tengelymetszet ($\hat{a}$) tesztelése

Teszteljük a hipotézist, hogy $H_0 : a = 0$ (a tengelymetszet nulla) és $H_1 : a \neq 0$.
A próbastatisztika képlete:

$$t = \frac{\hat{a}}{SE(\hat{a})}$$

ahol $SE(\hat{a})$, a tengelymetszet standard hibája, az alábbi képlettel számítható:

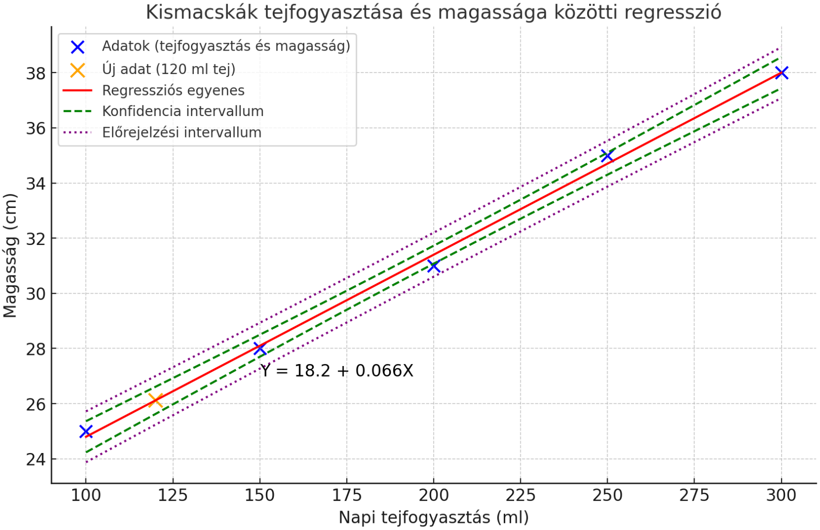
$$SE(\hat{a}) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}} = 0.7483$$

Számítás után a próbastatisztika:

$$t = \frac{18.2}{0.7483} = 24.32$$

A kritikus érték: $t_{3,0.975} = 3.182$. Mivel $t = 24.32 > 3.182$, elutasítjuk H_0-t, tehát a tengelymetszet szignifikánsan különbözik nullától.

Grafikon



Kismacskák tejfogyasztása és magassága közötti regresszió
Adatok (tejfogyasztás és magasság)
Új adat (120 ml tej)
Regressziós egyenes
Konfidencia intervallum
Előrejelzési intervallum
Y = 18.2 + 0.066X
Magasság (cm)
Napi tejfogyasztás (ml)
24
26
28
30
32
34
36
38
100
125
150
175
200
225
250
275
300

Kismacskák regressziós modellje - 1/1 rész

R-kód: Adatok és regressziós egyenlet

```
# Adatok
X <- c(100, 150, 200, 250, 300)
Y <- c(25, 28, 31, 35, 38)
# Lineáris regresszió
model <- lm(Y   X)
# Regressziós egyenlet együtthatók
a <- coef(model)[1] # Tengelymetszet
b <- coef(model)[2] # Meredekség
# Regressziós egyenlet kiíratása
cat("A regressziós egyenlet: Y = ", round(a, 2), " + ", round(b, 4), "X")
# A modell összefoglalása
summary(model)
```

Kismacskák regressziós modellje - 1/2 rész

Output:

```
A regressziós egyenlet: Y = 18.2 + 0.0660X
```

Modell összefoglalása:

Residuals:

Min	1Q	Median	3Q	Max
-0.6000	-0.5500	0.0000	0.5500	0.6000

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.2000	0.7483	24.32	0.0017**
X	0.0660	0.0030	22.11	0.0020**

```
---
Residual standard error: 0.6 on 3 degrees of freedom
Multiple R-squared: 0.9935, Adjusted R-squared: 0.9913
F-statistic: 488.9 on 1 and 3 DF, p-value: 0.002047
```

Kismacskák regressziós modellje - 2. rész

R-kód: Új kismacska becslése és konfidenciaintervallum

```
# Új megfigyelés
X_new <- 120 # Új tejfogyasztás (ml)
Y_new <- predict(model, newdata = data.frame(X = X_new))
# Konfidenciaintervallum számítása
conf_interval <- predict(model, newdata = data.frame(X = X_new), interval =
"confidence", level = 0.95)
# Output kiíratása
cat("A becsült magasság: ", round(Y_new, 2), "cm")
cat("Konfidenciaintervallum (95%): [", round(conf_interval[2], 2), ",",
round(conf_interval[3], 2), "] cm")
```

Output:

```
A becsült magasság: 26.12 cm
Konfidenciaintervallum (95%): [25.63, 26.61] cm
```

Értelmezés: Az új kismacska magasságának becsült értéke 26.12 cm. A 95%-os konfidenciaintervallum szerint a magasság 25.63 cm és 26.61 cm között várható.

Kismacskák regressziós modellje - 3/1. rész

R-kód: Előrejelzési intervallum és grafikon

```
# Előrejelzési intervallum számítása
pred_interval <- predict(model, newdata = data.frame(X = X_new), interval =
"prediction", level = 0.95)
# Output kiíratása
cat("Előrejelzési intervallum (95%): [", round(pred_interval[2], 2), ",",
round(pred_interval[3], 2), "] cm")
# Grafikon elkészítése
plot(X, Y, pch = 16, col = "blue", xlab = "Napi tejfogyasztás (ml)", ylab =
"Magasság (cm)", main = "Kismacskák regressziós modellje")
abline(model, col = "red", lwd = 2)
# Konfidenciaintervallum rajzolása
new_X <- seq(min(X), max(X), length.out = 100)
conf_interval <- predict(model, newdata = data.frame(X = new_X), interval =
"confidence")
lines(new_X, conf_interval[,2], col = "green", lty = 2)
lines(new_X, conf_interval[,3], col = "green", lty = 2)
# Előrejelzési intervallum rajzolása
pred_interval <- predict(model, newdata = data.frame(X = new_X), interval =
"prediction")
lines(new_X, pred_interval[,2], col = "purple", lty = 3)
lines(new_X, pred_interval[,3], col = "purple", lty = 3)
# Új adatpont hozzáadása
points(X_new, Y_new, pch = 16, col = "orange", cex = 2)
```

Kismacskák regressziós modellje - 3/2. rész

Output:

```
Előrejelzési intervallum (95%): [25.24, 27.00] cm
```

Értelmezés: Az előrejelzési intervallum szélesebb, figyelembe véve a megfigyelések közötti szóródást. A 95%-os előrejelzési intervallum 25.24 cm és 27.00 cm között van.