

MSc - Statisztika

Házi feladat

Kiss Dániel Márk

2023

# Contents

|                  |                   |               |
|------------------|-------------------|---------------|
| <b>Chapter 1</b> | <b>1. feladat</b> | <b>Page 2</b> |
| 1.1              | a) feladat        | 2             |
| 1.2              | b) feladat        | 3             |
| 1.3              | c) feladat        | 4             |
| 1.4              | d) feladat        | 4             |
| 1.5              | e) feladat        | 5             |
| <b>Chapter 2</b> | <b>2.feladat</b>  | <b>Page 6</b> |
| 2.1              | a) feladat        | 6             |
| <b>Chapter 3</b> | <b>3.feladat</b>  | <b>Page 7</b> |
| 3.1              | a) feladat        | 7             |
| 3.2              | b) feladat        | 7             |
| 3.3              | c) feladat        | 7             |

# Chapter 1

## 1. feladat

Az elmúlt évek kutatásai arra irányultak, hogy felmérjék a mosolygós emoji használatának hatását a digitális kommunikációban és a felhasználók boldogságszintjére. Az alábbi adatokat gyűjtötték össze: bead11.1.csv.

Lineáris regressziós modellt szeretnénk felírni, melyben az eredményváltozó a boldogságszint, míg a magyarázó változók az üzenet hossza és a mosolygós emoji száma.

### 1.1 a) feladat

Beccsüld meg és értelmezd a lineáris regresszió paramétereit, teszteld le, szignifikánsak-e a magyarázó változók! (5%-os szignifikanciaszinten)

Megodlás: A bead11.1.csv fájl négy oszlopot tartalmaz: "Sorszám," "Üzenet hossza," "Mosolygós emoji száma," és "Boldogságszint." A lineáris regresszió célja az, hogy a függő változót (pl. Boldogságszint) lineáris kapcsolatban álló magyarázó változókkal (pl. Üzenet hossza, Mosolygós emoji száma) modellezze.

A lineáris regresszió modellje általánosan a következő alakú:  $Y=B_0+B_1X_1+B_2X_2+E$

Python kód:

```
import pandas as pd
import statsmodels.api as sm
from scipy.stats import shapiro

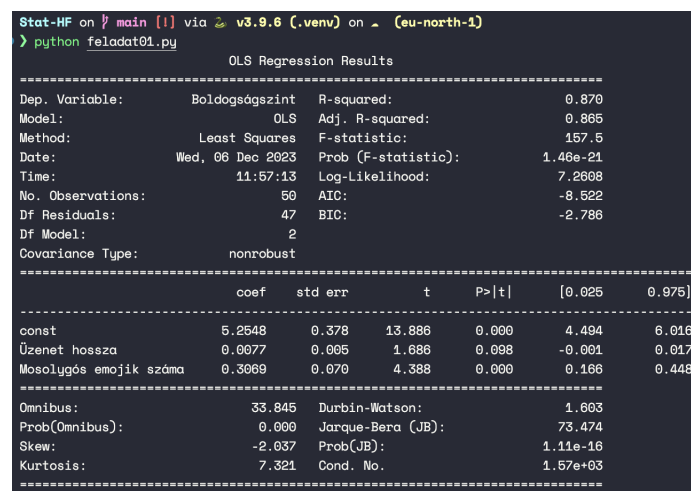
def main():
    """The main function of the program."""
    # 1.lepes: Adatok beolvasasa
    data = pd.read_csv("bead11.1.csv")
```

```
# 2.lepes: Magyarazo valtozok es a celvaltozo kivlasztasa
X = data[["Uzenet-hossza", "Mosolygos-emojik-szama"]]
X = sm.add_constant(X) # Allando hozzaadasa a magyarazo valtozokhoz
y = data["Boldogsagszint"] #Celvaltozo kivlasztasa

model = sm.OLS(y, X).fit()

# Regressziós modell parametereinek kiirasa
print(model.summary())
```

Kimenet értelmezése: Az alábbi kimenet a Figure 1.1-en látható. Az R-négyszet érték azt



```
Stat-HF on / main [1] via v3.9.6 (.venv) on (eu-north-1)
> python feladat01.py
```

| OLS Regression Results |                  |                     |          |       |        |        |
|------------------------|------------------|---------------------|----------|-------|--------|--------|
| Dep. Variable:         | Boldogsagszint   | R-squared:          | 0.870    |       |        |        |
| Model:                 | OLS              | Adj. R-squared:     | 0.865    |       |        |        |
| Method:                | Least Squares    | F-statistic:        | 157.5    |       |        |        |
| Date:                  | Wed, 06 Dec 2023 | Prob (F-statistic): | 1.46e-21 |       |        |        |
| Time:                  | 11:57:13         | Log-Likelihood:     | -7.2868  |       |        |        |
| No. Observations:      | 50               | AIC:                | -8.522   |       |        |        |
| Df Residuals:          | 47               | BIC:                | -2.786   |       |        |        |
| Df Model:              | 2                |                     |          |       |        |        |
| Covariance Type:       | nonrobust        |                     |          |       |        |        |
|                        | coef             | std err             | t        | P> t  | [0.025 | 0.975] |
| const                  | 5.2548           | 0.378               | 13.886   | 0.000 | 4.494  | 6.016  |
| Uzenet hossza          | 0.0077           | 0.005               | 1.686    | 0.098 | -0.001 | 0.017  |
| Mosolygos emojik száma | 0.3069           | 0.070               | 4.388    | 0.000 | 0.166  | 0.448  |
| Omnibus:               | 33.845           | Durbin-Watson:      | 1.603    |       |        |        |
| Prob(Omnibus):         | 0.000            | Jarque-Bera (JB):   | 73.474   |       |        |        |
| Skew:                  | -2.037           | Prob(JB):           | 1.11e-16 |       |        |        |
| Kurtosis:              | 7.321            | Cond. No.           | 1.57e+03 |       |        |        |

Figure 1.1: Feladat 1/a kimenet

mutatja, hogy a modell mennyire magyarázza a függő változó (Boldogságszint) változását. Az 0.870 érték azt jelenti, hogy a modell 87%-ban magyarázza a változást.  $P = \text{abs}(t)$  (szignifikanciaszint): Az egyes együtthatók (const, Üzenet hossza, Mosolygós emojik száma) szignifikanciaszintje. Az értékek alattuk a p-értékeket jelentik. Azok az együtthatók, amelyek p-értéke kevesebb, mint 0.05, szignifikánsak a 0.05 szignifikanciaszinten. Ebben a modellben mind a const (konstans), mind a Mosolygós emojik száma szignifikáns, mivel a p-értékük kisebb, mint 0.05. Összességében ez azt jelenti, hogy a modell jól teljesít a magyarázatokban, és mind a konstans, mind a Mosolygós emojik szám változói szignifikánsan kapcsolódnak a Boldogságszint változóhoz.

## 1.2 b) feladat

Határozd meg és értelmezd a többszörös determinációs együtthatót!

Megoldás:

A többszörös determinációs együttható (R-négyzet) azt mutatja, hogy a modell mennyire magyarázza a függő változó (Boldogságszint) változását. Az 0.870 érték azt jelenti, hogy a modell 87%-ban magyarázza a változást. Az R érték 0 és 1 közötti értéket vehet fel. Minél közelebb van az 1-hez, annál jobban magyarázza a modell a függő változó (Boldogságszint) változását. Az R-négyzet mellett fontos megjegyezni az "Adj. R-squared" értéket is (itt 0.865), amely korrigálja az R-négyzetet a magyarázó változók számára. Ez különösen fontos, ha több magyarázó változó van a modellben, mivel az R-négyzet hajlamos növekedni a változók számával anélkül, hogy ténylegesen javítaná a modell illeszkedését.

### 1.3 c) feladat

Teszteld a regressziós modell megbízhatóságát 5%-os szignifikanciaszinten!

Megoldás: A nullhipotézis az, hogy a modell nem szignifikáns azaz nincs összefüggés az emoji és a boldogságszint között. A nullhipotézis elutasításához a p-értéknek kisebbnek kell lennie, mint a szignifikanciaszint (5%). Az adott kimenetben a F-statistic értéke 157.5, és a hozzá tartozó p-érték a "Prob (F-statistic)" oszlopban található (1.46e-21). Ez az érték rendkívül kicsi, sok nagyságrenddel kisebb, mint 0.05 (5%-os szignifikanciaszint), így elvetjük a nullhipotézist (azaz elfogadjuk a modell szignifikanciáját). Ez azt jelenti, hogy a modell összességében szignifikánsan jól illeszkedik adatainkhoz.

### 1.4 d) feladat

Adj intervallumbecslést 95%-os megbízhatósággal paraméterekre!

Megoldás: Az "alpha=0.05" paraméter megadja a 95%-os megbízhatósági szintet. Python kód:

```
...  
# 4.lepes: Intervallumbecslés  
print(model.conf_int(alpha=0.05))  
...
```

Kimenet értelmezése:

Az intervallumok azt mutatják, hogy a konstans érték (const) becslési intervalluma 4.493509 és 6.016082 között van, az Üzenet hossza becslési intervalluma -0.001479 és 0.016810 között van, míg a Mosolygós emoji száma becslési intervalluma 0.166219 és 0.447623 között van. Ezek az intervallumok segíthetnek abban, hogy becsljük a paraméterek értékeinek megbízhatóságát és azt, mennyire pontosak a becslések.

|                       | 0         | 1        |
|-----------------------|-----------|----------|
| const                 | 4.493509  | 6.016082 |
| Üzenet hossza         | -0.001479 | 0.016810 |
| Mosolygós emoji száma | 0.166219  | 0.447623 |

## 1.5 e) feladat

Készíts előrejelzést az új üzenetek boldogságszintjére, ha az üzenet hossza 130 karakter, és a mosolygós emoji száma 3. Illetve adj ugyanerre 95%-os megbízhatóságú intervallumbecslést is.

Megoldás:

...

*# 5. feladat: Előrejelzés és intervallumbecslés*

```
uj_uzenet = pd.DataFrame(
    {"const": 1, "Uzenet_hossza": [130], "Mosolygos_emoji_szama": [3]}
)
predict_value = model.get_prediction(uj_uzenet).summary_frame()
print("Előrejelzés:", predict_value["mean"][0])
print(
    "95%-os megbízhatóságú intervallum:",
    predict_value["mean_ci_lower"][0],
    "_",
    predict_value["mean_ci_upper"][0],
)
```

Kimenet értelmezése: Előrejelzés: 7.172034019616269 95%-os megbízhatóságú intervallum: 7.09557490761663 - 7.248493131615907

Ez azt jelenti, hogy az új üzenetek boldogságszintje várhatóan körülbelül 7.2 lesz, és a 95%-os megbízhatóságú intervallum körülbelül 7.1 és 7.2 között lesz.

## Chapter 2

### 2.feladat

A következő kutatás arra irányult, hogy mérje a mosolygós emoji használatának hatását a kommunikációban különböző csoportokban. Az alábbi adatokat gyűjtötték össze: bead11.1.csv.

#### 2.1 a) feladat

Teszteld le, hogy van-e szignifikáns különbség a mosolygós emoji használatának gyakoriságában a különböző csoportokban ( $E = 0,05$  szignifikanciaszinten)!

Megoldás:

## Chapter 3

### 3.feladat

A bead11.3.csv file egy felmérés adatait mutatja a mosolygós emoji használatának változásáról az elmúlt években egy adott online fórumon.

#### 3.1 a) feladat

Készíts idősor diagramot az adatok alapján, majd számold ki a tapasztalati autokorrelációs és parciális autokorrelációs függvényeket.

Megoldás:

#### 3.2 b) feladat

Az adatok transzformációjával és a trend, valamint a szezonális komponensek kiszűrésével kísérletezve illessz különböző idősor modelleket. Teszteld az illeszkedést.

Megoldás:

#### 3.3 c) feladat

Készíts előrejelzést a következő hónapokra várható mosolygós emoji használatára.

Megoldás: