Projekt do předmětu MSP

December 11, 2022

Autor: Tereza Burianová, xburia28

[ ]:
```
import numpy as np
from scipy import stats
import openpyxl as px
import statsmodels.api as sm
import pandas as pd
```

1 Úkol 1

Pro výpočty v úkolu 1 je využit X^2 - test dobré shody.

Je třeba získat skutečné četnosti z odpovědí respondentů (f_{real_j}) a teoretické četnosti (f_{teor_j}) pomocí bodového odhadu p:

$p = x/n$, kde x je celkový počet respondentů, kterým vyhovuje zimní čas, a n je celkový počet respondentů.

Jednotlivé teoretické četnosti pak lze získat vynásobením počtu respondentů v dané skupině bodovým odhadem p:

$f_{teor_j} = p * n_j$

Testovací kritérium lze pak vypočítat následujícím způsobem:

$t = \sum_{j=1}^{m} \frac{(f_{real_j} - f_{teor_j})^2}{f_{teor_j}} \overset{as}{\sim} X^2(m-q-1)$, kde m je počet tříd a q je počet odhadovaných parametrů.

Doplněk kritického oboru lze vyjádřit následovně:

$\bar{W}_\alpha = \left\langle 0, X^2_{1-\alpha} \right\rangle$, kde $X^2_{1-\alpha}$ je kvantil Pearsonova rozdělení s m-q-1 stupni volnosti.

[ ]:
```
# 2 velka mesta, 2 mala mesta, 3 obce, 1 okoli studenta
resp = np.array([1327, 915, 681, 587, 284, 176, 215, 34])
zimni = np.array([510, 324, 302, 257, 147, 66, 87, 15])
letni = np.array([352, 284, 185, 178, 87, 58, 65, 8])
stridat = np.array([257, 178, 124, 78, 44, 33, 31, 4])
beznazoru = np.array([208, 129, 70, 74, 6, 19, 32, 7])

resp_sk = np.array([np.sum(resp[:2]), np.sum(resp[2:4]), np.sum(resp[4:7])])
```

```
zimni_sk = np.array([np.sum(zimni[:2]), np.sum(zimni[2:4]), np.sum(zimni[4:7])])
letni_sk = np.array([np.sum(letni[:2]), np.sum(letni[2:4]), np.sum(letni[4:7])])
stridat_sk = np.array([np.sum(stridat[:2]), np.sum(stridat[2:4]), np.
 ↪sum(stridat[4:7])])
beznazoru_sk = np.array([np.sum(beznazoru[:2]), np.sum(beznazoru[2:4]), np.
 ↪sum(beznazoru[4:7])])
```

[ ]:
```
def chisq(respondenti, testovane, odhad, st_volnosti):
    testovane_teor = respondenti * p
    print("Skutečné četnosti:", testovane)
    print("Teoretické četnosti: ", testovane_teor)
    t = np.sum(np.square(testovane - testovane_teor)/testovane_teor)
    print("Testovací kritérium: ", t)
    krit_obor = stats.chi2.ppf(0.95, df=st_volnosti)
    print("Doplněk kritického oboru: < 0,", krit_obor, ">")
```

a) H_0: V městech, obcích a v okolí studenta je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

H_A: V městech, obcích a v okolí studenta není stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Při ověřování následujících třech hypotéz (a), b) a c)) se pracovalo s 8 skupinami. Hodnota stupně volnosti bude 6, neboť je prováděn bodový odhad (tedy 8-1-1).

[ ]:
```
p = np.sum(zimni)/np.sum(resp)
chisq(resp, zimni, p, 6)
```

```
Skutečné četnosti: [510 324 302 257 147  66  87  15]
Teoretické četnosti:  [537.21640199 370.42427115 275.6928182  237.63830292
114.9732164
  71.25100735  87.03958284  13.76439915]
Testovací kritérium:  20.704110374775837
Doplněk kritického oboru: < 0, 12.591587243743977 >
```

$20,704 \notin \langle 0; 12,592 \rangle$, tedy H_0 zamítáme.

b) H_0: V městech, obcích a v okolí studenta je stejné procentuální zastoupení obyvatel, co preferují letní čas.

H_A: V městech, obcích a v okolí studenta není stejné procentuální zastoupení obyvatel, co preferují letní čas.

[ ]:
```
p = np.sum(letni)/np.sum(resp)
chisq(resp, letni, p, 6)
```

```
Skutečné četnosti: [352 284 185 178  87  58  65   8]
Teoretické četnosti:  [382.78241289 263.938137   196.4392036  169.32424745
81.92178241
  50.76842854  62.01825077   9.80753733]
```

```
Testovací kritérium:  6.932364791415857
Doplněk kritického oboru: < 0, 12.591587243743977 >
```

$6,932 \in \langle 0; 12,592 \rangle$, tedy H_0 nezamítáme.

c) H_0: V městech, obcích a v okolí studenta je stejné procentuální zastoupení obyvatel, co preferují střídání času.

H_A: V městech, obcích a v okolí studenta není stejné procentuální zastoupení obyvatel, co preferují střídání času.

[ ]:
```
p = np.sum(stridat)/np.sum(resp)
chisq(resp, stridat, p, 6)
```

```
Skutečné četnosti: [257 178 124  78  44  33  31   4]
Teoretické četnosti:  [235.58260251 162.44015169 120.89808011 104.21023939
50.4185826
  31.2453188   38.16899739   6.03602749]
Testovací kritérium:  13.058303417150736
Doplněk kritického oboru: < 0, 12.591587243743977 >
```

$13,058 \notin \langle 0; 12,592 \rangle$, tedy H_0 zamítáme.

d) H_0: U větších měst, menších měst a obcí je stejné procentuální zastoupení obyvatel, co preferují zimní čas.

H_A: U větších měst, menších měst a obcí není stejné procentuální zastoupení obyvatel, co preferují zimní čas.

Při ověřování následujících dvou hypotéz (d), e)) se pracuje se 3 skupinami, sloučenými podle větších měst, menších měst a obcí. Hodnota stupně volnosti bude tedy 1, neboť je prováděn bodový odhad (tedy 3-1-1).

[ ]:
```
p = np.sum(zimni_sk)/np.sum(resp_sk)
chisq(resp_sk, zimni_sk, p, 1)
```

```
Skutečné četnosti: [834 559 300]
Teoretické četnosti:  [906.97873357 512.9567503  273.06451613]
Testovací kritérium:  12.661948651569508
Doplněk kritického oboru: < 0, 3.841458820694124 >
```

$12,66 \notin \langle 0; 3,84 \rangle$, tedy H_0 zamítáme.

e) H_0: U větších měst, menších měst a obcí je stejné procentuální zastoupení nerozhodnutelných obyvatel.

H_A: U větších měst, menších měst a obcí není stejné procentuální zastoupení nerozhodnutelných obyvatel.

[ ]:
```
p = np.sum(beznazoru_sk)/np.sum(resp_sk)
chisq(resp_sk, beznazoru_sk, p, 1)
```

```
Skutečné četnosti: [337 144  57]
Teoretické četnosti:  [288.21887694 163.00692951  86.77419355]
```

```
Testovací kritérium:  20.688664757394136
Doplněk kritického oboru: < 0, 3.841458820694124 >
```

$20,69 \notin \langle 0; 3,84 \rangle$, tedy H_0 zamítáme.

f) Na základě odpovědí z okolí studenta zkuste určit z dat, zda student prováděl výzkum ve větším městě, menším městě nebo v obci. Porovnejte výsledek se skutečností a okomentujte.

[ ]:
```
np.set_printoptions(suppress=True)
def chisq_f(sk_index):
    p_zimni = np.sum(zimni_sk[sk_index])/np.sum(resp_sk[sk_index])
    p_letni = np.sum(letni_sk[sk_index])/np.sum(resp_sk[sk_index])
    p_stridani = np.sum(stridat_sk[sk_index])/np.sum(resp_sk[sk_index])
    p_beznazoru = np.sum(beznazoru_sk[sk_index])/np.sum(resp_sk[sk_index])
    vyzkum = [zimni[7], letni[7], stridat[7], beznazoru[7]]
    celkem = np.sum(vyzkum)
    vyzkum_teor = [celkem*p_zimni, celkem*p_letni, celkem*p_stridani,␣
 ↪celkem*p_beznazoru]
    print(stats.chisquare(f_obs = vyzkum, f_exp = vyzkum_teor, ddof=2))

krit_obor = stats.chi2.ppf(0.95, df=2)
print("Doplněk kritického oboru: < 0,", krit_obor, ">")
print("H0: Student prováděl výzkum ve větším městě.")
chisq_f(0)
print("H0: Student prováděl výzkum v menším městě.")
chisq_f(1)
print("H0: Student prováděl výzkum v obci.")
chisq_f(2)
```

```
Doplněk kritického oboru: < 0, 5.991464547107979 >
H0: Student prováděl výzkum ve větším městě.
Power_divergenceResult(statistic=2.438784599437811, pvalue=0.11836790712014167)
H0: Student prováděl výzkum v menším městě.
Power_divergenceResult(statistic=3.230669566985398, pvalue=0.07227113427356965)
H0: Student prováděl výzkum v obci.
Power_divergenceResult(statistic=6.947865988500661, pvalue=0.008391928065668765)
```

Na základě p-values lze zjistit, že hypotézy o provedení výzkumu ve větším nebo menším městě zamítnuty nebyly, naopak hypotéza o provedení výzkumu v obci byla zamítnuta. Tyto výsledky jsou srovnatelné s realitou, neboť výzkum byl prováděn v menším městě i větším městě, s převahou obyvatel z většího města. Dá se tedy vyvodit, že pozorované a dodané odpovědi souhlasí.

2 Úkol 2

a) Určení vhodného modelu.

Pro vytvoření modelu a další práci s ním byla využita knihovna *statsmodels*.

Daný model $Z = \beta_1 + \beta_2 X + \beta_3 Y + \beta_4 X^2 + \beta_5 Y^2 + \beta_6 X * Y$ může být zjednodušen až na model $Z = \beta_1$. Je vhodné odstranit nulové parametry pro zjednodušení modelu. Ty mohou být určeny

například hodnotou P>|t|, která značí p-value pro hypotézu, že daný parametr je nulový. Pokud je tedy tato hodnota vyšší, než hodnota $\alpha = 0,05$, nezamítá se hypotéza, že je parametr nulový, a může být potenciálně odstraněn. Vždy musí být přihlíženo na vhodnost modelu, která může být určena pomocí koeficientu determinace R^2.

[ ]:
```
W = px.load_workbook('data.xlsx')
p = W['data']
dataX = np.asarray([ p['A%s'%i].value for i in range(5,75) ])
dataY = np.asarray([ p['B%s'%i].value for i in range(5,75) ])
dataZ = np.asarray([ p['D%s'%i].value for i in range(5,75) ])

F = np.column_stack((dataX, dataY, dataX**2, dataY**2, dataX*dataY))
F = sm.add_constant(F)
model = sm.OLS(dataZ, F).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.981
Model:                            OLS   Adj. R-squared:                  0.980
Method:                 Least Squares   F-statistic:                     661.5
Date:                Sun, 11 Dec 2022   Prob (F-statistic):           1.14e-53
Time:                        16:16:35   Log-Likelihood:                -242.05
No. Observations:                  70   AIC:                             496.1
Df Residuals:                      64   BIC:                             509.6
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.9131      3.846     -0.757      0.452     -10.596       4.770
x1             0.6598      0.600      1.100      0.276      -0.539       1.859
x2            -0.0397      1.133     -0.035      0.972      -2.302       2.223
x3             0.4663      0.027     17.424      0.000       0.413       0.520
x4            -0.0662      0.100     -0.664      0.509      -0.266       0.133
x5            -1.0241      0.045    -22.688      0.000      -1.114      -0.934
==============================================================================
Omnibus:                        0.359   Durbin-Watson:                   1.712
Prob(Omnibus):                  0.836   Jarque-Bera (JB):                0.391
Skew:                          -0.162   Prob(JB):                        0.822
Kurtosis:                       2.830   Cond. No.                         839.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Podle hodnot P>|t| lze rozpoznat, že parametry x1, x2 a x4 lze odstranit. Hodnota R^2 je bez

odstranění parametrů 0,981.

```
[ ]: F = np.column_stack((dataX**2, dataX*dataY))
     F = sm.add_constant(F)
     submodel = sm.OLS(dataZ, F).fit()
     print(submodel.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.979
Method:                 Least Squares   F-statistic:                     1610.
Date:                Sun, 11 Dec 2022   Prob (F-statistic):           2.29e-57
Time:                        16:16:35   Log-Likelihood:                -244.54
No. Observations:                  70   AIC:                             495.1
Df Residuals:                      67   BIC:                             501.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4852      1.478     -1.681      0.097      -5.436       0.466
x1             0.5065      0.009     55.012      0.000       0.488       0.525
x2            -1.0657      0.024    -44.353      0.000      -1.114      -1.018
==============================================================================
Omnibus:                        1.755   Durbin-Watson:                   1.618
Prob(Omnibus):                  0.416   Jarque-Bera (JB):                1.312
Skew:                          -0.332   Prob(JB):                        0.519
Kurtosis:                       3.094   Cond. No.                         307.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Model byl postupně zjednodušen na funkci $Z = \beta_1 + \beta_4 X^2 + \beta_6 X * Y$. Rozdíl mezi původní a novou hodnotou R^2, která je nyní 0,980, není zásadní, tedy vytvořený submodel je možno použít.

b) Odhady regresních parametrů metodou nejmenších čtverců a jejich 95% intervaly spolehlivosti.

Tyto hodnoty byly zjištěny z výše uvedeného shrnutí.

	[0.025	0.975]	Koeficient
const	-5.436	0.466	-2.4852
x1	0.488	0.525	0.5065
x2	-1.114	-1.018	-1.0657

c) Nestranně odhadněte rozptyl závislé proměnné.

Nestranný odhad rozptylu závislé proměnné Z lze určit jako $S^2 = \frac{RSS}{n-2}$, kde RSS vyjadřuje reziduální součet čtverců. Tuto hodnotu lze z výše uvedeného modelu získat pomocí *mse_resid*, tedy $S^2 = 66,2064$.

[ ]:
```
submodel.mse_resid
```

[ ]:
```
66.20643157748601
```

d) Vhodným testem zjistěte, že vámi zvolené dva regresní parametry jsou současně nulové.

H_0: parametry x1 = x2 = 0

H_A: některý z parametrů není roven 0

Pro ověření hypotézy nulovosti parametrů je vhodný f-test. Hodnota $F \overset{as}{\sim} F_{1-\alpha}(k1, k2)$, $k1 = I - 1$ a $k2 = n - I$, kde n je počet realizací a I je počet skupin (koeficientů). V tomto případě se jedná o hodnotu $F_{0,95}(1, 68)$. Tedy $F = 0,6249 \in \langle 0, 3, 9819 \rangle$, hypotéza H_0 se tak na hladině významnosti 0,05 nezamítá.

[ ]:
```
print(model.f_test("x1=x2=0"))
print(stats.f.ppf(q=0.95, dfn=1, dfd=68))
```

```
<F test: F=0.6248911439742278, p=0.5385537138776201, df_denom=64, df_num=2>
3.9818962563017606
```

e) Vhodným testem zjistěte, že vámi zvolené dva regresní parametry jsou současně nulové.

H_0: parametry x1 = x2

H_A: parametry x1 != x2

Pro ověření této hypotézy byl využit t-test. Lze pozorovat, že $t \notin \langle 1,511; 1,633 \rangle$, tedy hypotéza H_0 se tak na hladině významnosti 0,05 zamítá.

[ ]:
```
print(submodel.t_test("x1=x2"))
```

```
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0             1.5722      0.030     51.566      0.000       1.511       1.633
==============================================================================
```