RE: 07/18 DM課上午場

洪子軒
Sent: Monday, July 18, 2016 4:09 PM
To: 洪子軒

http://bse.nchu.edu.tw/new_page_535.htm
http://statsmodels.sourceforge.net/devel/example_formulas.html

X變數用愈多，R square 會越大，造成解釋能力虛化
所以要改用 → Adjusted R Square

--
降低維度（避免模型失效、視覺化、屬性不獨立造成重複解釋）

屬性重複
屬性不相關

方法一、Feature Selection
方法二、Feature Extraction
＞pca 主成份分析 Principle Component Analysis
＞svd Singular Value Decomposition
＞mds Multi-dimensional scaling

--
機率密度 cdf pdf

empirical normal probability 常態分配
http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.norm.html

poisson distribution 離散分配
二項分配

--
對不同迴歸模型作 anova 測試比較

http://statsmodels.sourceforge.net/devel/examples/generated/example_interactions.html

```
from statsmodels.stats.anova import anova_lm
table1 = anova_lm(lm, interX_lm)
print table1
```

p 值 >= 0.05 代表檢定的對象彼此相同（不拒絕虛無假設）
p 值 < 0.05 有顯著差異（拒絕虛無假設—兩者相同）

--
用來衡量迴歸模型複雜度的方式

參數用得越多，模型越複雜，成本就越高
aic https://en.wikipedia.org/wiki/Akaike_information_criterion
bic https://en.wikipedia.org/wiki/Bayesian_information_criterion
說明 https://read01.com/zPjKQK.html

--
統計共線性檢定（X變數夠不夠獨立）
變異數影響因子（Variance Inflation Factor，VIF）

http://statsmodels.sourceforge.net/devel/_modules/statsmodels/stats/outliers_influence.html#variance_inflation_factor

--
X vs. Y

・類別/類別：卡方檢定
・類別/數值：ANOVA
例：abc 教學法影響成績
・數值/類別：均值T檢定（先做變異數F檢定）
例：平均身高對於性別有無顯著；好公司和壞公司的平均固定資產周轉率有無不同
・數值/數值：相關係數

--
http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#example-classification-plot-classifier-comparison-py

```
from sklearn.preprocessing import StandardScaler
```

```
X = StandardScaler().fit_transform(X)
```

http://docs.scipy.org/doc/scipy/reference/stats.html
雙尾、單尾

--
洪子軒 Tzu-Hsuan Hung
中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號
--

From: 洪子軒
Sent: Monday, July 18, 2016 11:57 AM
To: 洪子軒
Subject: RE: 07/18 DM課上午場

--
洪子軒 Tzu-Hsuan Hung
中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號
--

From: 洪子軒
Sent: Monday, July 18, 2016 11:49 AM
To: 洪子軒
Subject: 07/18 DM課上午場

```
>>> x = pandas.Series(np.random.randn(10))
>>> stats.skew(x)
-0.17644348972413657
>>> x.skew()
-0.20923623968879457
>>> stats.skew(x, bias=False)
-0.2092362396887948
>>> stats.kurtosis(x)
0.6362620964462327
>>> x.kurtosis()
2.0891062062174464
>>> stats.kurtosis(x, bias=False)
2.089106206217446
```

--
常態分配
http://stackoverflow.com/questions/13865596/quantile-quantile-plot-using-scipy

```
import numpy as np
import pylab
import scipy.stats as stats

measurements = np.random.normal(loc = 20, scale = 5, size=100)
stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```

--
shapiro-wilk normality test 常態分配檢定測試
http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.shapiro.html

--
雙群樣本 Avg 平均數檢定 = T test (相減) ；var 變異數檢定 F test (相除)

```
>>> import pandas as pd
>>> import scipy.stats
>>> import numpy as np
>>> df_a = pd.read_clipboard()
>>> df_b = df_a + np.random.randn(5, 7)
>>> df_c = df_a + np.random.randn(5, 7)
>>> t_b, p_b = scipy.stats.ttest_ind(df_a.dropna(axis=0), df_b.dropna(axis=0))
>>> t_b, p_c = scipy.stats.ttest_ind(df_a.dropna(axis=0), df_c.dropna(axis=0))
>>> pd.DataFrame([p_b, p_c], columns = df_a.columns, index = ['df_b', 'df_c'])
      VSPD1_perc  VSPD2_perc  VSPD3_perc  VSPD4_perc  VSPD5_perc  VSPD6_perc  \
df_b    0.425286    0.987956    0.644236    0.552244    0.432640    0.624528
df_c    0.947182    0.911384    0.189283    0.828780    0.697709    0.166956

      VSPD7_perc
df_b    0.546648
df_c    0.206950
```

p 值<0.05 對立 (不相等) 假設成立

--
ANOVA 檢定：多群平均值是否彼此相等
http://www.marsja.se/four-ways-to-conduct-one-way-anovas-using-python/

```
# compute one-way ANOVA P value
from scipy import stats

f_val, p_val = stats.f_oneway(treatment1, treatment2, treatment3)

print "One-way ANOVA P =", p_val

One-way ANOVA P = 0.381509481874
```

If P > 0.05, we can claim with high confidence that the means of the results of all three experiments are not significantly different.

--
卡方檢定/比例檢定：離散、數目count 檢定

http://codereview.stackexchange.com/questions/96761/chi-square-independence-test-for-two-pandas-df-columns

```
def chi_square_of_df_cols(df,col1,col2):
    return scs.chi2_contingency([
        [
            len(df[(df[col1] == cat) & (df[col2] == cat2)])
            for cat2 in range(int(df[col1].min()), int(df[col1].max()) + 1)
        ]
        for cat in range(int(df[col2].min()), int(df[col2].max()) + 1)
    ])
```

http://stackoverflow.com/questions/11725115/p-value-from-chi-sq-test-statistic-in-python

--
Outlier 離群值 mahalanobis distance (欄位屬性彼此不獨立)

http://stackoverflow.com/questions/29817090/is-there-a-python-equivalent-to-the-mahalanobis-function-in-r-if-not-how-can

```
from scipy.spatial.distance import mahalanobis
import scipy as sp
import pandas as pd

x = pd.read_csv('IrisData.csv')
x = x.ix[:,1:]

Sx = x.cov().values
Sx = sp.linalg.inv(Sx)

mean = x.mean().values

def mahalanobisR(X,meanCol,IC):
    m = []
    for i in range(X.shape[0]):
        m.append(mahalanobis(X.ix[i,:],meanCol,IC) ** 2)
    return(m)

mR = mahalanobisR(x,mean,Sx)
```

```
stats.chi2.cdf()
```

To calculate probability of null hypothesis given chisquared sum, and degrees of freedom you can also call `chisqprob`:

```
>>> from scipy.stats import chisqprob
>>> chisqprob(3.84, 1)
0.050043521248705189
```

--

http://stackoverflow.com/questions/19991445/run-an-ols-regression-with-pandas-data-frame

R-squared / adjusted R-squared

```
>>> import pandas as pd
>>> import statsmodels.formula.api as sm
>>> df = pd.DataFrame({"A": [10,20,30,40,50], "B": [20, 30, 10, 40, 50], "C": [32, 234, 23, 23,
>>> result = sm.ols(formula="A ~ B + C", data=df).fit()
>>> print result.params
Intercept    14.952480
B             0.401182
C             0.000352
dtype: float64
>>> print result.summary()
                            OLS Regression Results
==============================================================================
Dep. Variable:                      A   R-squared:                       0.579
Model:                            OLS   Adj. R-squared:                  0.158
Method:                 Least Squares   F-statistic:                     1.375
Date:                Thu, 14 Nov 2013   Prob (F-statistic):              0.421
Time:                        20:04:30   Log-Likelihood:                -18.178
No. Observations:                   5   AIC:                             42.36
Df Residuals:                       2   BIC:                             41.19
Df Model:                           2
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     14.9525     17.764      0.842      0.489       -61.481    91.386
B              0.4012      0.650      0.617      0.600        -2.394     3.197
C              0.0004      0.001      0.650      0.583        -0.002     0.003
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.061
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.498
Skew:                          -0.123   Prob(JB):                        0.780
Kurtosis:                       1.474   Cond. No.                     5.21e+04
```

--

洪子軒 Tzu-Hsuan Hung
中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號

--