cancer_deathrate_prediction_OLS_regression

April 24, 2018

```
In [1]: %matplotlib inline
        import pandas as pd
        import numpy as np
        from sklearn import linear_model
        from sklearn import ensemble
        from sklearn.model_selection import train_test_split
        import matplotlib.pyplot as plt
```

```
In [2]: df = pd.read_csv('cancer_reg.csv')
        df.head()
```

```
Out[2]:    avgAnnCount  avgDeathsPerYear  TARGET_deathRate  incidenceRate  medIncome  \
        0       1397.0               469             164.9          489.8      61898
        1        173.0                70             161.3          411.6      48127
        2        102.0                50             174.7          349.7      49348
        3        427.0               202             194.8          430.4      44243
        4         57.0                26             144.4          350.1      49955

           popEst2015  povertyPercent  studyPerCap            binnedInc  MedianAge  \
        0      260131            11.2   499.748204   (61494.5, 125635]       39.3
        1       43269            18.6    23.111234  (48021.6, 51046.4]       33.0
        2       21026            14.6    47.560164  (48021.6, 51046.4]       45.0
        3       75882            17.1   342.637253    (42724.4, 45201]       42.8
        4       10321            12.5     0.000000  (48021.6, 51046.4]       48.3

             ...      PctPrivateCoverageAlone  PctEmpPrivCoverage PctPublicCoverage  \
        0    ...                          NaN                41.6              32.9
        1    ...                         53.8                43.6              31.1
        2    ...                         43.5                34.9              42.1
        3    ...                         40.3                35.0              45.3
        4    ...                         43.9                35.1              44.0

           PctPublicCoverageAlone   PctWhite  PctBlack  PctAsian  PctOtherRace  \
        0                    14.0  81.780529  2.594728  4.821857      1.843479
        1                    15.3  89.228509  0.969102  2.246233      3.741352
        2                    21.1  90.922190  0.739673  0.465898      2.747358
        3                    25.0  91.744686  0.782626  1.161359      1.362643
```

```
4                     22.7  94.104024  0.270192  0.665830      0.492135

   PctMarriedHouseholds  BirthRate
0             52.856076   6.118831
1             45.372500   4.333096
2             54.444868   3.729488
3             51.021514   4.603841
4             54.027460   6.796657

[5 rows x 34 columns]
```

```
In [3]: # missing value handling
        df['PctSomeCol18_24'] = df['PctSomeCol18_24'].fillna(df['PctSomeCol18_24'].mean())
        df['PctEmployed16_Over'] = df['PctEmployed16_Over'].fillna(df['PctEmployed16_Over'].mean
        df['PctPrivateCoverageAlone'] = df['PctPrivateCoverageAlone'].fillna(df['PctPrivateCover
```

```
In [4]: y = df['TARGET_deathRate']
        y.head()
```

```
Out[4]: 0    164.9
        1    161.3
        2    174.7
        3    194.8
        4    144.4
        Name: TARGET_deathRate, dtype: float64
```

```
In [5]: x = df
        x = x.drop('TARGET_deathRate', axis=1)
        x = x.drop('binnedInc', axis=1) #removed binnedInc field
        x.head()
        x = pd.get_dummies(x)
        x.head()
```

```
Out[5]:    avgAnnCount  avgDeathsPerYear  incidenceRate  medIncome  popEst2015  \
        0       1397.0               469          489.8      61898      260131
        1        173.0                70          411.6      48127       43269
        2        102.0                50          349.7      49348       21026
        3        427.0               202          430.4      44243       75882
        4         57.0                26          350.1      49955       10321

           povertyPercent  studyPerCap  MedianAge  MedianAgeMale  MedianAgeFemale  \
        0            11.2   499.748204       39.3           36.9             41.7
        1            18.6    23.111234       33.0           32.2             33.7
        2            14.6    47.560164       45.0           44.0             45.8
        3            17.1   342.637253       42.8           42.2             43.4
        4            12.5     0.000000       48.3           47.8             48.9

                          ...                Geography_York County, Pennsylvania  \
        0                 ...                                                  0
```

```
1               ...                                                  0
2               ...                                                  0
3               ...                                                  0
4               ...                                                  0

   Geography_York County, South Carolina  Geography_York County, Virginia  \
0                                      0                                0
1                                      0                                0
2                                      0                                0
3                                      0                                0
4                                      0                                0

   Geography_Young County, Texas  Geography_Yuba County, California  \
0                              0                                  0
1                              0                                  0
2                              0                                  0
3                              0                                  0
4                              0                                  0

   Geography_Yukon-Koyukuk Census Area, Alaska  \
0                                            0
1                                            0
2                                            0
3                                            0
4                                            0

   Geography_Yuma County, Arizona  Geography_Yuma County, Colorado  \
0                               0                                0
1                               0                                0
2                               0                                0
3                               0                                0
4                               0                                0

   Geography_Zapata County, Texas  Geography_Zavala County, Texas
0                               0                              0
1                               0                              0
2                               0                              0
3                               0                              0
4                               0                              0

[5 rows x 3078 columns]
```

```
In [6]: train_x,test_x,train_y,test_y = train_test_split(x,y)
```

```
In [7]: lr = linear_model.LinearRegression()
        lr.fit(train_x,train_y)
```

```
Out[7]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

```
In [8]: lr.intercept_
```

```
Out[8]: 166.75087262224966
```

```
In [9]: lr.coef_
```

```
Out[9]: array([-2.86465051e-03,  1.60135840e-02,  1.92088818e-01, ...,
                1.48800406e+01,  0.00000000e+00, -4.57450314e+01])
```

```
In [10]: print('new data score: ',lr.score(test_x,test_y), 'same data score: ',lr.score(train_x,
```

```
('new data score: ', 0.4270123712613325, 'same data score: ', 1.0)
```