MSDM 5054

Assignment 1

Problem 1

1. **Wrong.** The better a function fits the data, the more it fits the noise, called overfitting. Thus, overfitting happens when the model is too flexible to fit both the significant data and the useless noise. An overfitting model may perform quite well on the training set but poorly on the test set. Zero training error is not necessary.
2. **Decrease.** In the KNN algorithm, when K is small, the model is quite sensitive to changes in the local structure of the dataset, leading to significant variations in performance across different data, which means higher variance. As K increases, the model becomes smoother, the bias in predictions increases, but the variance decreases.
3. $\int(1-\max q_j(x))p(x)dx$
4. **The sum of residuals should be zero** in multiple linear regression. In linear regression, to minimize the Residual Sum of Squares, the first derivative is set to zero, resulting in the sum of residuals being zero.
5. **R square can be used to compare two models with same size.** R square means we always choose the model with smaller SSerror which can maximize the R square. However, the complex one with more parameters always has larger R square due to its smaller bias if there are two models with different size. For two same size models, R square works well.
6. **Linear Regression**

   Pros: performs well with less data and easy to inference with relatively high interpretability.

   Cons: its form could be totally wrong; less flexibility.

   KNN Regression

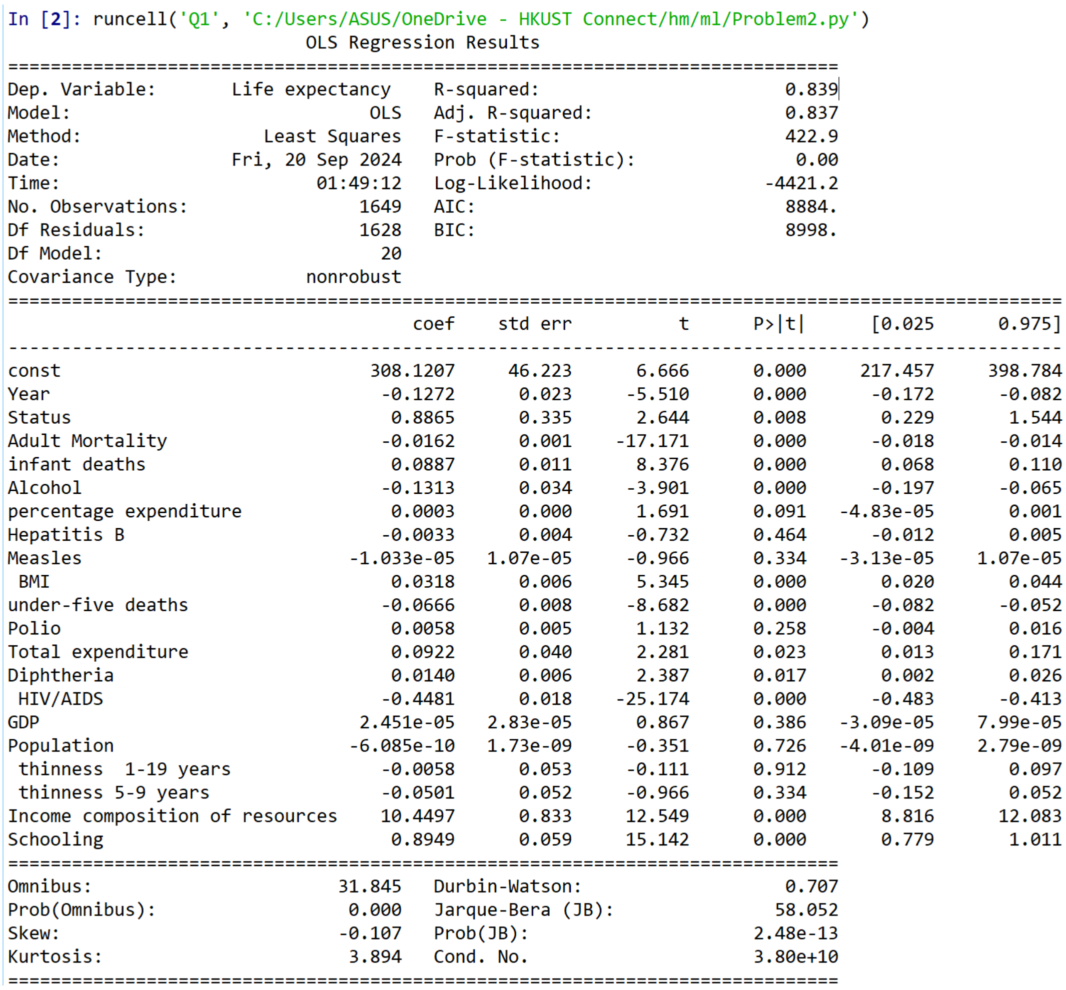
   Pros: fits wider range of shapes of function as a non-parametric method.

   Cons: it needs a lot of data to work well compared with Linear Regression.

   It would be hard to do inference work due to its complicate form.

Problem 2

1. Model summary:



```
In [2]: runcell('Q1', 'C:/Users/ASUS/OneDrive - HKUST Connect/hm/ml/Problem2.py')
                            OLS Regression Results                            
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.839
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                     422.9
Date:                Fri, 20 Sep 2024   Prob (F-statistic):               0.00
Time:                        01:49:12   Log-Likelihood:                -4421.2
No. Observations:                1649   AIC:                             8884.
Df Residuals:                    1628   BIC:                             8998.
Df Model:                          20                                         
Covariance Type:            nonrobust                                         
=====================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                             308.1207     46.223      6.666      0.000     217.457     398.784
Year                               -0.1272      0.023     -5.510      0.000      -0.172      -0.082
Status                              0.8865      0.335      2.644      0.008       0.229       1.544
Adult Mortality                    -0.0162      0.001    -17.171      0.000      -0.018      -0.014
infant deaths                       0.0887      0.011      8.376      0.000       0.068       0.110
Alcohol                            -0.1313      0.034     -3.901      0.000      -0.197      -0.065
percentage expenditure              0.0003      0.000      1.691      0.091   -4.83e-05       0.001
Hepatitis B                        -0.0033      0.004     -0.732      0.464      -0.012       0.005
Measles                         -1.033e-05   1.07e-05     -0.966      0.334   -3.13e-05    1.07e-05
 BMI                                0.0318      0.006      5.345      0.000       0.020       0.044
under-five deaths                  -0.0666      0.008     -8.682      0.000      -0.082      -0.052
Polio                               0.0058      0.005      1.132      0.258      -0.004       0.016
Total expenditure                   0.0922      0.040      2.281      0.023       0.013       0.171
Diphtheria                          0.0140      0.006      2.387      0.017       0.002       0.026
 HIV/AIDS                          -0.4481      0.018    -25.174      0.000      -0.483      -0.413
GDP                              2.451e-05   2.83e-05      0.867      0.386   -3.09e-05    7.99e-05
Population                      -6.085e-10   1.73e-09     -0.351      0.726   -4.01e-09    2.79e-09
 thinness  1-19 years              -0.0058      0.053     -0.111      0.912      -0.109       0.097
 thinness 5-9 years                -0.0501      0.052     -0.966      0.334      -0.152       0.052
Income composition of resources    10.4497      0.833     12.549      0.000       8.816      12.083
Schooling                           0.8949      0.059     15.142      0.000       0.779       1.011
==============================================================================
Omnibus:                       31.845   Durbin-Watson:                   0.707
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               58.052
Skew:                          -0.107   Prob(JB):                     2.48e-13
Kurtosis:                       3.894   Cond. No.                     3.80e+10
==============================================================================
```

From this report, we can roughly think ***Year, Adult Mortality, infant death, Alcohol, BMI, under-five deaths, HIV/AIDS, Income composition of resources and Schooling*** are significant variables affecting the life expectancy. However, there may be some strong multicollinearity problems because we choose too many covariates without further selection, and it is hard to scale the range of all these variables.

2. 95% confidence interval for the coefficients of “Adult Mortality”: - 0.0406 ~ - 0.0352

   95% confidence interval for the coefficients of “HIV/AIDS”: - 0.4862 ~ - 0.3734

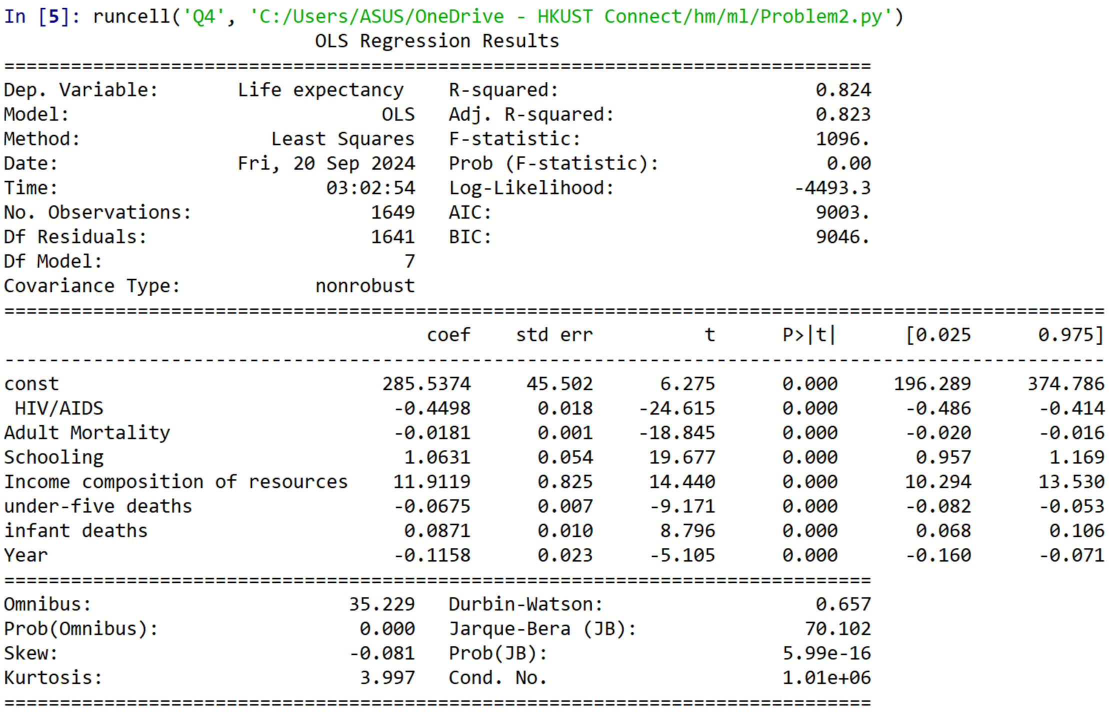
   Although the coefficients are significant negative (95% confidence intervals are lower than zero), we can not be sure that these predictors really have negative impact on the life expectancy. It is because we ignore the impact of other variables when we construct the two factors simple model. For example, some of these ignored variables may have complex interaction effect with the two factors and lead to the final impact on the life expectancy.

3. The 97% confidence interval for the coefficient of “Schooling”: 2.288345 ~ 2.580960

   The 97% confidence interval for the coefficient of “Alcohol”: -0.264407 ~ -0.061395

   The interval of “Schooling” is above zero, which means that “Schooling” has a positive impact on the life expectancy with 97% confidence. The interval of “Alcohol” is slightly lower than zero, which means that “Alcohol” has a negative impact to some degree on the life expectancy with 97% confidence.

4. The top seven most influential predictors are HIV/AIDS, Adult Mortality, Schooling, Income composition of resources, under-five deaths, infant deaths, and Year. The summary is showed as below.



```
In [5]: runcell('Q4', 'C:/Users/ASUS/OneDrive - HKUST Connect/hm/ml/Problem2.py')
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.824
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     1096.
Date:                Fri, 20 Sep 2024   Prob (F-statistic):               0.00
Time:                        03:02:54   Log-Likelihood:                -4493.3
No. Observations:                1649   AIC:                             9003.
Df Residuals:                    1641   BIC:                             9046.
Df Model:                           7
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                             285.5374     45.502      6.275      0.000     196.289     374.786
 HIV/AIDS                          -0.4498      0.018    -24.615      0.000      -0.486      -0.414
Adult Mortality                    -0.0181      0.001    -18.845      0.000      -0.020      -0.016
Schooling                           1.0631      0.054     19.677      0.000       0.957       1.169
Income composition of resources    11.9119      0.825     14.440      0.000      10.294      13.530
under-five deaths                  -0.0675      0.007     -9.171      0.000      -0.082      -0.053
infant deaths                       0.0871      0.010      8.796      0.000       0.068       0.106
Year                               -0.1158      0.023     -5.105      0.000      -0.160      -0.071
==============================================================================
Omnibus:                       35.229   Durbin-Watson:                   0.657
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               70.102
Skew:                          -0.081   Prob(JB):                     5.99e-16
Kurtosis:                       3.997   Cond. No.                     1.01e+06
==============================================================================
```

5. Predict the Life Expectancy as around 88.38

   The 99% Confidence Interval: 85.99 ~ 90.77

6. Full Model AIC: 8884.483951849954

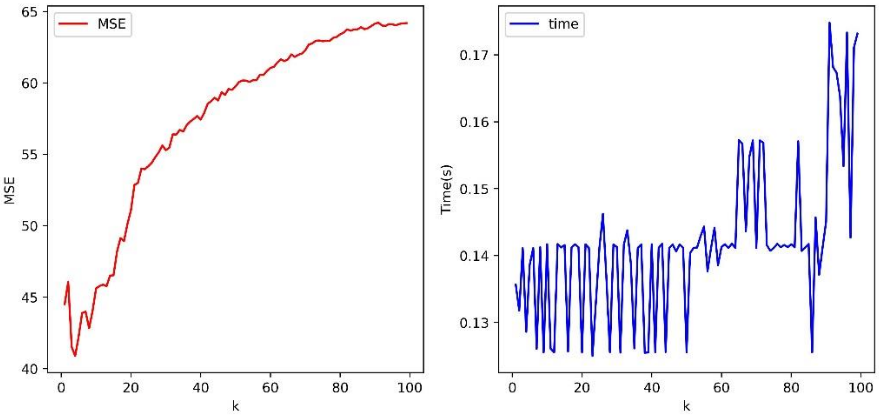
   Small Model AIC: 9002.562179694152

   The model with smallest AIC is preferred. Thus, it seems that Full Model is preferred than the smaller model.

Problem 3

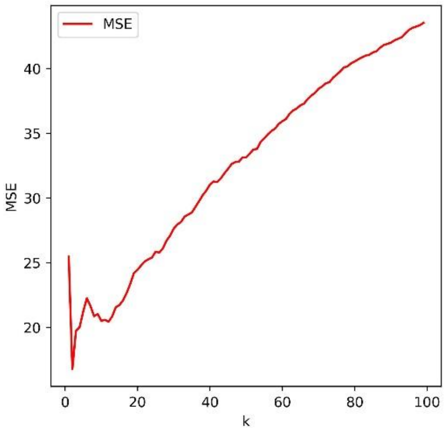
1. Test MSE reach the minimum at around 40.88 when K=4. The best value of K is 4.

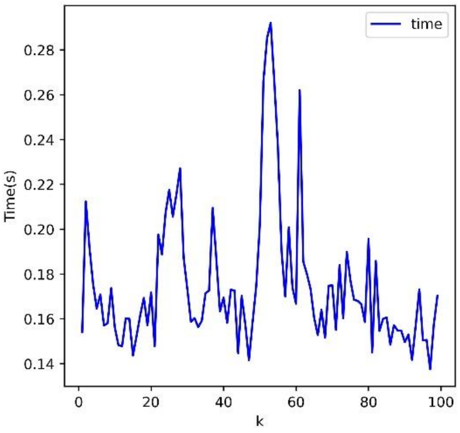
K (see ‘k_range’ in code)	MSE (see ‘MeanSE’)	Time
1	44.51708661417323	0.1268603801727295
2	46.05590551181103	0.1264653205871582
3	41.52323709536307	0.14185070991516113
4	40.88791830708661	0.1414954662322998
5	42.24112440944882	0.1416645050048828
6	43.889407261592304	0.12514352798461914
7	43.985068295034544	0.1412184238433838
8	42.83032357283464	0.12839269638061523
9	44.04344026441139	0.13933968544006348
… (k range from 1 to 100)	…	…



2. After standardize the predictor variables (excluding the response variable) in both training and test data, test MSE reach the minimum at around 16.77 when K=2. The best value of K is 2.

K (see ‘k_range’ in code)	MSE (see ‘MeanSE’)	Time
1	25.46929133858267	0.1542191505432129
2	16.777578740157477	0.2123730182647705
3	19.73487314085739	0.19198060035705566
4	20.019940944881885	0.17559599876403809
5	21.226831496062992	0.16461682319641113
6	22.255448381452318	0.17090654373168945
7	21.654661738711233	0.15705180168151855
8	20.868348917322834	0.15802001953125
9	21.036780402449686	0.17373251914978027
… (k range from 1 to 100)	…	…





3. Standardization significantly reduced the test MSE, improving the predictive performance of KNN regression, but there was no noticeable change in the model’s runtime.

Note: copilot tool is used to Copilot is used to query the necessary third-party library interfaces and some functional functions (such as the model.pvalues.sort_values() used in Problem 2).