output

July 5, 2023

1 Data sets for testing the program

1.1 1. Predicting Medical Insurance Costs

1.1.1 Summary

From a data set that compiles information on peoples' medical history we implement a linear regression model that attempts to predict the insurance costs of patients.

Data Set Description (source)

- `age`: age of primary beneficiary
- `sex`: insurance contractor gender, female, male
- `bmi`: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
- `objective` index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- `children`: Number of children covered by health insurance / Number of dependents
- `smoker`: Smoking
- `region`: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- `charges`: Individual medical costs billed by health insurance

1.2 2. Predicting Real Estate Value In The Suburbs Of Boston

1.2.1 Summary

Predict the price of real estate based on different characterizing factors.

Data Set Description (source)

- `CRIM` - per capita crime rate by town.
- `ZN` - proportion of residential land zoned for lots over 25,000 sq.ft.
- `INDUS` - proportion of non-retail business acres per town.
- `CHAS` - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- `NOX` - nitrogen oxides concentration (parts per 10 million).
- `RM` - average number of rooms per dwelling.
- `AGE` - proportion of owner-occupied units built prior to 1940.
- `DIS` - weighted mean of distances to five Boston employment centres.
- `RAD` - index of accessibility to radial highways.

- TAX - full-value property-tax rate per $10,000.
- PTRATIO - pupil-teacher ratio by town.
- LSTAT - lower status of the population (percent).
- MEDV - median value of owner-occupied homes in $1000s.

```
Control script: ctrl_insurance.py


#### Initial Variables:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
None


Data Viewer 1:

    age     sex     bmi  children smoker     region      charges
0    19  female  27.900         0    yes  southwest  16884.92400
1    18    male  33.770         1     no  southeast   1725.55230
2    28    male  33.000         3     no  southeast   4449.46200
3    33    male  22.705         0     no  northwest  21984.47061
4    32    male  28.880         0     no  northwest   3866.85520
5    31  female  25.740         0     no  southeast   3756.62160
6    46  female  33.440         1     no  southeast   8240.58960
7    37  female  27.740         3     no  northwest   7281.50560
8    37    male  29.830         2     no  northeast   6406.41070
9    60  female  25.840         0     no  northwest  28923.13692
10   25    male  26.220         0     no  northeast   2721.32080



Variable Description Before Data Processing:

                age          bmi     children       charges
count  1338.000000  1338.000000  1338.000000   1338.000000
```

```
mean     39.207025    30.663397     1.094918  13270.422265
std      14.049960     6.098187     1.205493  12110.011237
min      18.000000    15.960000     0.000000   1121.873900
25%      27.000000    26.296250     0.000000   4740.287150
50%      39.000000    30.400000     1.000000   9382.033000
75%      51.000000    34.693750     2.000000  16639.912515
max      64.000000    53.130000     5.000000  63770.428010


Missing Values (NAs) Per Column:

age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64


Missing values (NAs) After Replacement/Removal:

age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64


#### Variables After Transformation:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   age          1338 non-null   int64
 1   bmi          1338 non-null   float64
 2   children     1338 non-null   int64
 3   region       1338 non-null   object
 4   charges      1338 non-null   float64
 5   sex_d        1338 non-null   int64
 6   smoker_d     1338 non-null   int64
 7   log_charges  1338 non-null   float64
```

```
dtypes: float64(3), int64(4), object(1)
memory usage: 83.8+ KB
None
```

Data Viewer 2:

	age	bmi	children	region	charges	sex_d	smoker_d	log_charges
0	19	27.900	0	southwest	16884.92400	0	0	9.734176
1	18	33.770	1	southeast	1725.55230	1	1	7.453302
2	28	33.000	3	southeast	4449.46200	1	1	8.400538
3	33	22.705	0	northwest	21984.47061	1	1	9.998092
4	32	28.880	0	northwest	3866.85520	1	1	8.260197
5	31	25.740	0	southeast	3756.62160	0	1	8.231275
6	46	33.440	1	southeast	8240.58960	0	1	9.016827
7	37	27.740	3	northwest	7281.50560	0	1	8.893093
8	37	29.830	2	northeast	6406.41070	1	1	8.765054
9	60	25.840	0	northwest	28923.13692	0	1	10.272397
10	25	26.220	0	northeast	2721.32080	1	1	7.908873

Checking Non-numerical Variables:

- Non-numeric variables in the main data frame:
  - region

- Non-numeric variables dropped.

Pearson's Correlations

	log_charges	smoker_d	sex_d	charges	children	bmi
age	0.528	0.025	-0.021	0.299	0.042	0.109
bmi	0.133	-0.004	0.046	0.198	0.013	-
children	0.161	-0.008	0.017	0.068	-	-
charges	0.893	-0.787	0.057	-	-	-
sex_d	0.006	-0.076	-	-	-	-
smoker_d	-0.666	-	-	-	-	-

Analysis Of Variance Inflation Factor:

Features group: 1

sex_d	1.000435
age	1.000435

```
Name: VIF, dtype: float64

-----------------------------------

Features group: 2

sex_d    1.002838
age      1.012775
bmi      1.014516
Name: VIF, dtype: float64

-----------------------------------

Features group: 3

age         1.000988
sex_d       1.006202
smoker_d    1.006394
Name: VIF, dtype: float64

-----------------------------------

Features group: 4

children    1.002242
smoker_d    1.006457
sex_d       1.008878
bmi         1.014578
age         1.015129
Name: VIF, dtype: float64

-----------------------------------


#### Feature Selection Algorithms

## Univariate Selection (Select k Best):

Container index: 1

Parameters: {'target': 'charges', 'k_vars': 4, 'criterion': 'f_regression'}

Variables selected:

        - 'age'
        - 'bmi'
        - 'children'
        - 'smoker_d'
```

```
-----------------------------------

Container index: 2

Parameters: {'target': 'log_charges', 'k_vars': 4, 'criterion': 'f_regression'}

Variables selected:

        - 'age'
        - 'bmi'
        - 'children'
        - 'smoker_d'

--------------------------------------------------------------------------------
----


## Unique Combinations Of Explanatory Variables Derived From The Feature
Selection Stage:

- Target: 'charges'
    - ['age', 'bmi', 'children', 'smoker_d']

- Target: 'log_charges'
    - ['age', 'bmi', 'children', 'smoker_d']


Total: 2


#### Regression Results


## Summary:

Regression Nr: 1:
    - Selection strategy: univariate
    - Target: 'charges'
    - Explanatory Variables: ['age', 'bmi', 'children', 'smoker_d']

Regression Nr: 2:
    - Selection strategy: univariate
    - Target: 'log_charges'
    - Explanatory Variables: ['age', 'bmi', 'children', 'smoker_d']

Regression Nr: 3:
    - Selection strategy: manually selected
```

```
    - Target: 'charges'
    - Explanatory Variables: ['age', 'sex_d']

Regression Nr: 4:
    - Selection strategy: manually selected
    - Target: 'charges'
    - Explanatory Variables: ['age', 'sex_d', 'bmi']

Regression Nr: 5:
    - Selection strategy: manually selected
    - Target: 'charges'
    - Explanatory Variables: ['age', 'sex_d', 'smoker_d']

Regression Nr: 6:
    - Selection strategy: manually selected
    - Target: 'charges'
    - Explanatory Variables: ['age', 'sex_d', 'bmi', 'children', 'smoker_d']




### Regression number: 1

Target variable (Y): 'charges'

Explanatory Variables:

    - x1: 'age'
    - x2: 'bmi'
    - x3: 'children'
    - x4: 'smoker_d'

                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.741
Dependent Variable: y                AIC:                21706.1732
Date:               2023-07-05 16:47 BIC:                21731.0503
No. Observations:   1070             Log-Likelihood:     -10848.
Df Model:           4                F-statistic:        764.4
Df Residuals:       1065             Prob (F-statistic): 3.77e-311
R-squared:          0.742            Scale:              3.7641e+07
------------------------------------------------------------------
         Coef.     Std.Err.    t     P>|t|     [0.025      0.975]
------------------------------------------------------------------
const  11580.5841 1116.1004  10.3759 0.0000   9390.5787  13770.5895
x1       257.5558   13.5998  18.9383 0.0000    230.8704    284.2412
x2       320.1445   30.8758  10.3688 0.0000    259.5602    380.7288
x3       556.2747  158.0091   3.5205 0.0004    246.2302    866.3193
```

```
x4    -23597.6939  465.1911 -50.7269 0.0000 -24510.4891 -22684.8987
------------------------------------------------------------------
Omnibus:              221.025       Durbin-Watson:          2.073
Prob(Omnibus):        0.000         Jarque-Bera (JB):       455.726
Skew:                 1.174         Prob(JB):               0.000
Kurtosis:             5.170         Condition No.:          310
==================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 32623008.47
RMSE: 5711.66


### Regression number: 2

Target variable (Y): 'log_charges'

Explanatory Variables:

    - x1: 'age'
    - x2: 'bmi'
    - x3: 'children'
    - x4: 'smoker_d'

                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.745
Dependent Variable: y                AIC:                1369.0198
Date:               2023-07-05 16:47 BIC:                1393.8969
No. Observations:   1070             Log-Likelihood:     -679.51
Df Model:           4                F-statistic:        783.7
Df Residuals:       1065             Prob (F-statistic): 1.98e-315
R-squared:          0.746            Scale:              0.20949
------------------------------------------------------------------
           Coef.    Std.Err.      t       P>|t|     [0.025    0.975]
------------------------------------------------------------------
const      8.5427     0.0833   102.5989   0.0000    8.3794    8.7061
x1         0.0340     0.0010    33.5202   0.0000    0.0320    0.0360
x2         0.0107     0.0023     4.6477   0.0000    0.0062    0.0152
x3         0.1016     0.0118     8.6210   0.0000    0.0785    0.1248
x4        -1.5172     0.0347   -43.7170   0.0000   -1.5853   -1.4491
------------------------------------------------------------------
Omnibus:              347.799      Durbin-Watson:         2.020
Prob(Omnibus):        0.000        Jarque-Bera (JB):      1143.341
```

```
Skew:                 1.589        Prob(JB):              0.000
Kurtosis:             6.943        Condition No.:         310
=================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the
errors is correctly specified.
```

Error measurement:

MSE: 0.18
RMSE: 0.42

Regression number: 3

Target variable (Y): 'charges'

Explanatory Variables:

- x1: 'age'
- x2: 'sex_d'

```
                  Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.088
Dependent Variable: y                AIC:                23049.7374
Date:               2023-07-05 16:47 BIC:                23064.6636
No. Observations:   1070             Log-Likelihood:     -11522.
Df Model:           2                F-statistic:        52.66
Df Residuals:       1067             Prob (F-statistic): 1.55e-22
R-squared:          0.090            Scale:              1.3237e+08
-----------------------------------------------------------------
            Coef.     Std.Err.     t     P>|t|    [0.025     0.975]
-----------------------------------------------------------------
const     2605.5835  1128.9424   2.3080  0.0212  390.3843  4820.7826
x1         255.4248    25.2778  10.1047  0.0000  205.8251   305.0246
x2        1462.8520   703.7863   2.0785  0.0379   81.8898  2843.8142
-----------------------------------------------------------------
Omnibus:              292.892       Durbin-Watson:          1.933
Prob(Omnibus):        0.000         Jarque-Bera (JB):       581.679
Skew:                 1.642         Prob(JB):               0.000
Kurtosis:             4.506         Condition No.:          139
=================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
```

Error measurement:

```
MSE: 128284398.24
RMSE: 11326.27


### Regression number: 4

Target variable (Y): 'charges'

Explanatory Variables:

    - x1: 'age'
    - x2: 'sex_d'
    - x3: 'bmi'

                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.113
Dependent Variable: y                AIC:                23021.0051
Date:               2023-07-05 16:47 BIC:                23040.9068
No. Observations:   1070             Log-Likelihood:     -11507.
Df Model:           3                F-statistic:        46.45
Df Residuals:       1066             Prob (F-statistic): 3.26e-28
R-squared:          0.116            Scale:              1.2875e+08
------------------------------------------------------------------
           Coef.     Std.Err.    t    P>|t|     [0.025     0.975]
------------------------------------------------------------------
const    -6373.0901 1958.3127 -3.2544 0.0012 -10215.6752 -2530.5049
x1         238.2345   25.1191  9.4842 0.0000    188.9459   287.5230
x2        1199.9390  695.6781  1.7248 0.0848   -165.1149  2564.9929
x3         318.9589   57.2301  5.5733 0.0000    206.6624   431.2554
------------------------------------------------------------------
Omnibus:               233.907       Durbin-Watson:          1.933
Prob(Omnibus):         0.000         Jarque-Bera (JB):       403.213
Skew:                  1.434         Prob(JB):               0.000
Kurtosis:              3.907         Condition No.:          292
==================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 126155971.63
RMSE: 11231.92


### Regression number: 5
```

Target variable (Y): 'charges'

Explanatory Variables:

    - x1: 'age'
    - x2: 'sex_d'
    - x3: 'smoker_d'

```
                  Results: Ordinary least squares
==============================================================================
Model:              OLS              Adj. R-squared:     0.712
Dependent Variable: y                AIC:                21819.0882
Date:               2023-07-05 16:47 BIC:                21838.9899
No. Observations:   1070             Log-Likelihood:     -10906.
Df Model:           3                F-statistic:        880.1
Df Residuals:       1066             Prob (F-statistic): 7.72e-288
R-squared:          0.712            Scale:              4.1869e+07
------------------------------------------------------------------------------
          Coef.    Std.Err.    t     P>|t|     [0.025      0.975]
------------------------------------------------------------------------------
const   21255.4996 744.2216  28.5607 0.0000  19795.1941  22715.8051
x1        277.1770  14.2235  19.4872 0.0000    249.2677    305.0862
x2         42.4621 396.9149   0.1070 0.9148   -736.3612    821.2854
x3     -23630.4702 491.9384 -48.0354 0.0000 -24595.7476 -22665.1927
------------------------------------------------------------------------------
Omnibus:              198.150       Durbin-Watson:          2.072
Prob(Omnibus):        0.000         Jarque-Bera (JB):       413.935
Skew:                 1.053         Prob(JB):               0.000
Kurtosis:             5.202         Condition No.:          173
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
```

Error measurement:

MSE: 36000763.69
RMSE: 6000.06

Regression number: 6

Target variable (Y): 'charges'

Explanatory Variables:

    - x1: 'age'

```
    - x2: 'sex_d'
    - x3: 'bmi'
    - x4: 'children'
    - x5: 'smoker_d'

                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.741
Dependent Variable: y                AIC:                21707.6711
Date:               2023-07-05 16:47 BIC:                21737.5236
No. Observations:   1070             Log-Likelihood:     -10848.
Df Model:           5                F-statistic:        611.4
Df Residuals:       1064             Prob (F-statistic): 8.65e-310
R-squared:          0.742            Scale:              3.7659e+07
------------------------------------------------------------------
         Coef.     Std.Err.    t     P>|t|     [0.025      0.975]
------------------------------------------------------------------
const  11699.2207 1128.9147  10.3632 0.0000   9484.0686  13914.3727
x1       257.2126   13.6116  18.8965 0.0000    230.5039    283.9212
x2      -266.7664  377.4882  -0.7067 0.4799  -1007.4724    473.9395
x3       321.6202   30.9536  10.3904 0.0000    260.8832    382.3572
x4       559.8364  158.1266   3.5404 0.0004    249.5610    870.1119
x5    -23622.1141  466.5819 -50.6280 0.0000 -24537.6392 -22706.5890
------------------------------------------------------------------
Omnibus:              220.123       Durbin-Watson:          2.075
Prob(Omnibus):        0.000         Jarque-Bera (JB):       451.903
Skew:                 1.172         Prob(JB):               0.000
Kurtosis:             5.156         Condition No.:          314
==================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 32669542.5
RMSE: 5715.73


** [No more experiments] **


[End Of Report]

[NbConvertApp] Making directory
../project_13/exported_pdf/notebook_version_new_version.pdf
[NbConvertApp] Converting notebook ../project_13/output.ipynb to pdf
[NbConvertApp] Writing 40065 bytes to notebook.tex
```

```
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 45093 bytes to
../project_13/exported_pdf/notebook_version_new_version.pdf/output.pdf
```