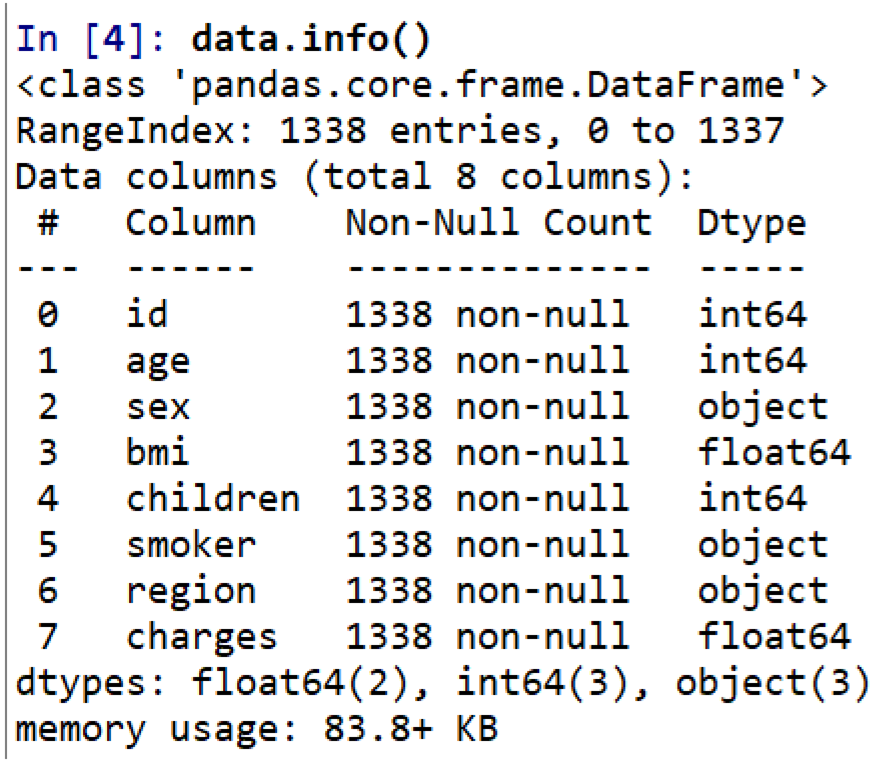
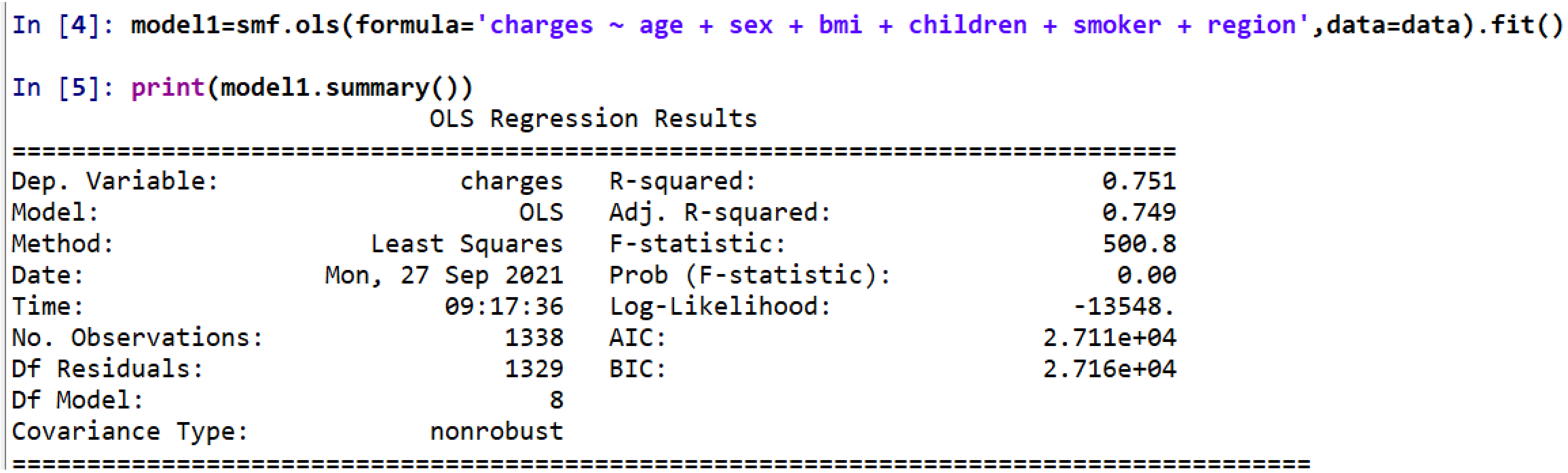
smf.ols vs sklean

Data: insurance



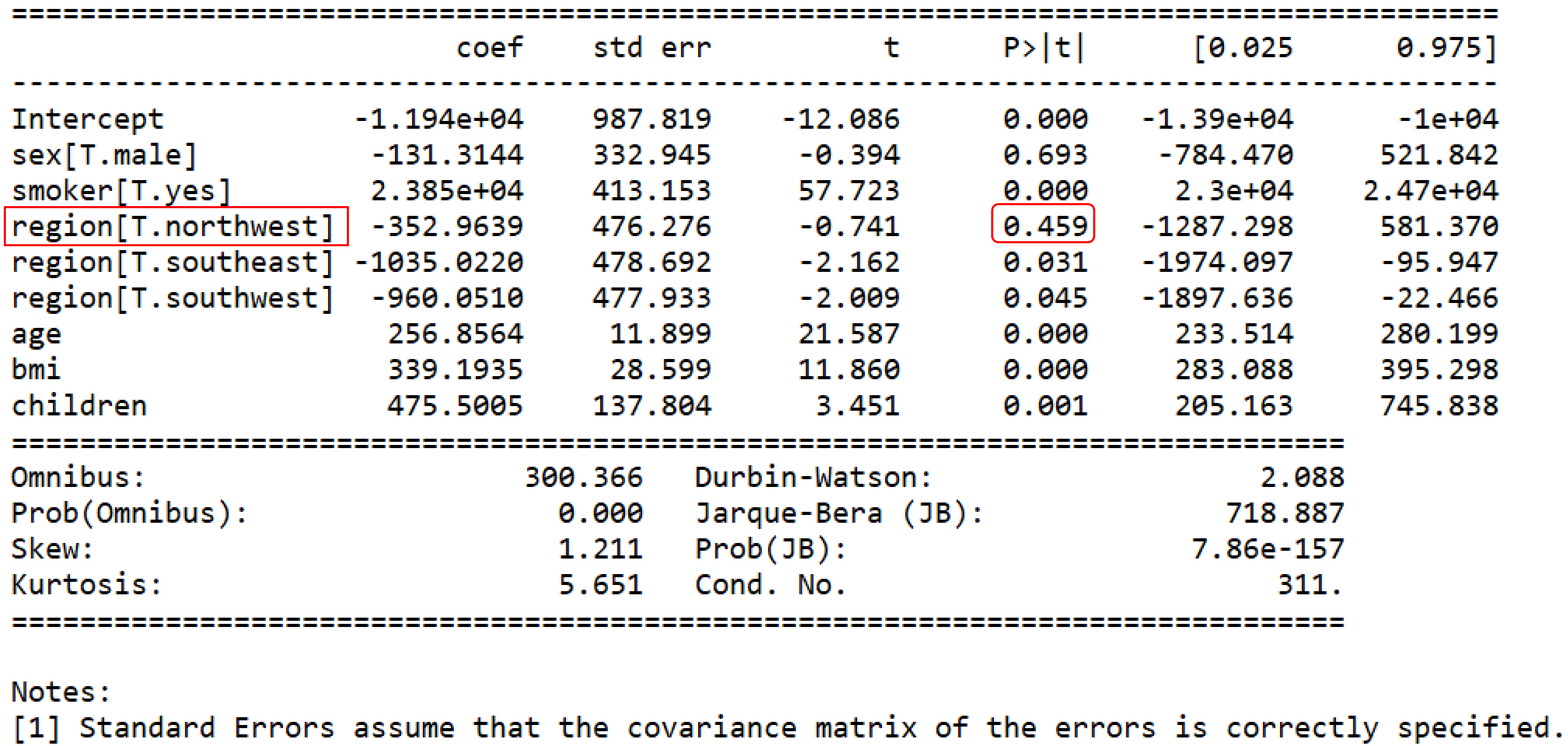
```
In [4]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   id        1338 non-null   int64  
 1   age       1338 non-null   int64  
 2   sex       1338 non-null   object 
 3   bmi       1338 non-null   float64
 4   children  1338 non-null   int64  
 5   smoker    1338 non-null   object 
 6   region    1338 non-null   object 
 7   charges   1338 non-null   float64
dtypes: float64(2), int64(3), object(3)
memory usage: 83.8+ KB
```

1st model



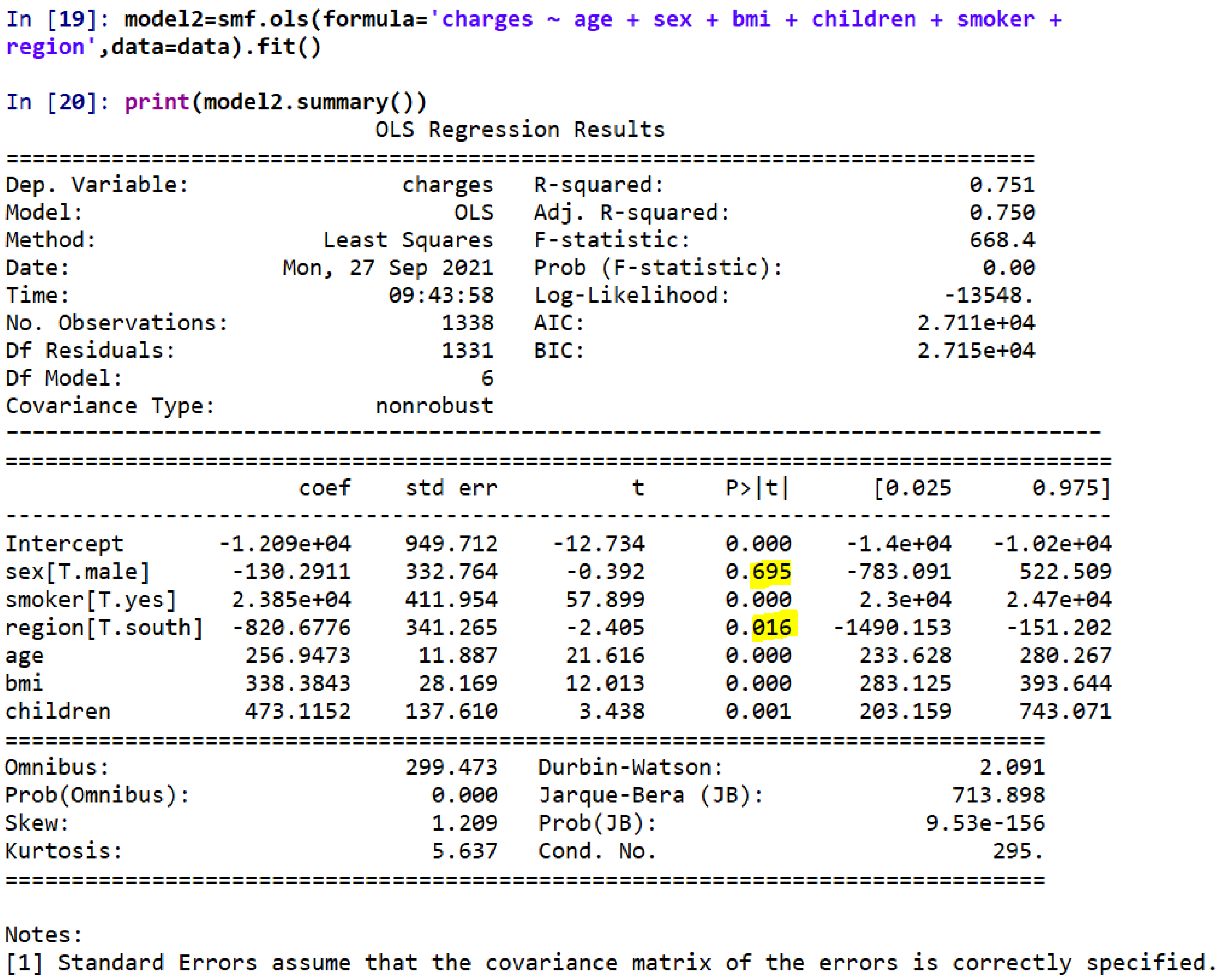
```
In [4]: model1=smf.ols(formula='charges ~ age + sex + bmi + children + smoker + region',data=data).fit()

In [5]: print(model1.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     500.8
Date:                Mon, 27 Sep 2021   Prob (F-statistic):               0.00
Time:                        09:17:36   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1329   BIC:                         2.716e+04
Df Model:                           8
Covariance Type:            nonrobust
=======================================================================================
```

1st model



```
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept           -1.194e+04    987.819    -12.086      0.000   -1.39e+04      -1e+04
sex[T.male]          -131.3144    332.945     -0.394      0.693    -784.470     521.842
smoker[T.yes]        2.385e+04    413.153     57.723      0.000     2.3e+04    2.47e+04
region[T.northwest]  -352.9639    476.276     -0.741      0.459   -1287.298     581.370
region[T.southeast] -1035.0220    478.692     -2.162      0.031   -1974.097     -95.947
region[T.southwest]  -960.0510    477.933     -2.009      0.045   -1897.636     -22.466
age                   256.8564     11.899     21.587      0.000     233.514     280.199
bmi                   339.1935     28.599     11.860      0.000     283.088     395.298
children              475.5005    137.804      3.451      0.001     205.163     745.838
==============================================================================
Omnibus:                      300.366   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.887
Skew:                           1.211   Prob(JB):                    7.86e-157
Kurtosis:                       5.651   Cond. No.                         311.
==============================================================================

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

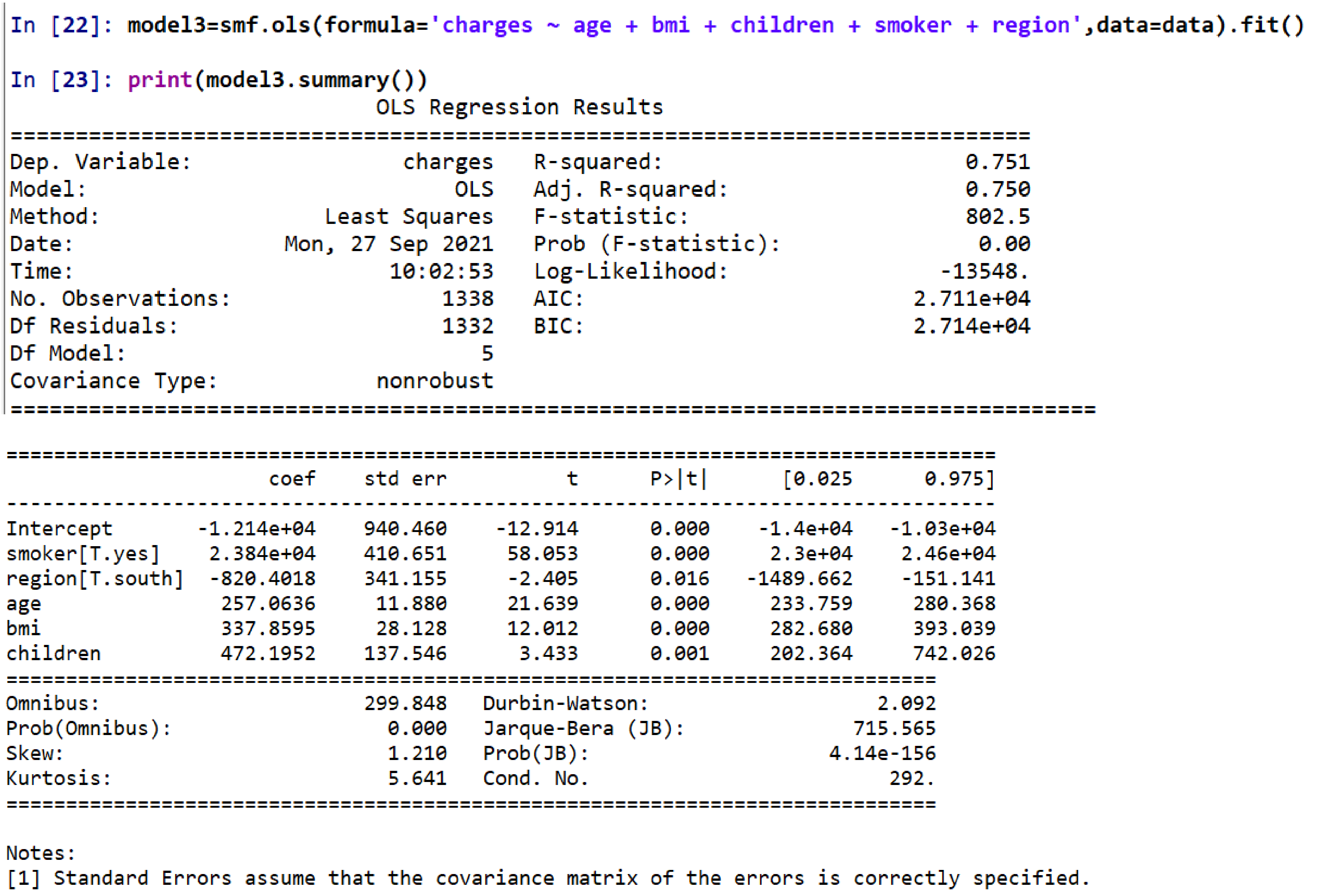
After region into 2 categories



```
In [19]: model2=smf.ols(formula='charges ~ age + sex + bmi + children + smoker +
region',data=data).fit()

In [20]: print(model2.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     668.4
Date:                Mon, 27 Sep 2021   Prob (F-statistic):               0.00
Time:                        09:43:58   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1331   BIC:                         2.715e+04
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept      -1.209e+04    949.712    -12.734      0.000    -1.4e+04   -1.02e+04
sex[T.male]     -130.2911    332.764     -0.392      0.695    -783.091     522.509
smoker[T.yes]   2.385e+04    411.954     57.899      0.000     2.3e+04    2.47e+04
region[T.south] -820.6776    341.265     -2.405      0.016   -1490.153    -151.202
age              256.9473     11.887     21.616      0.000     233.628     280.267
bmi              338.3843     28.169     12.013      0.000     283.125     393.644
children         473.1152    137.610      3.438      0.001     203.159     743.071
==============================================================================
Omnibus:                      299.473   Durbin-Watson:                   2.091
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              713.898
Skew:                           1.209   Prob(JB):                    9.53e-156
Kurtosis:                       5.637   Cond. No.                         295.
==============================================================================

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

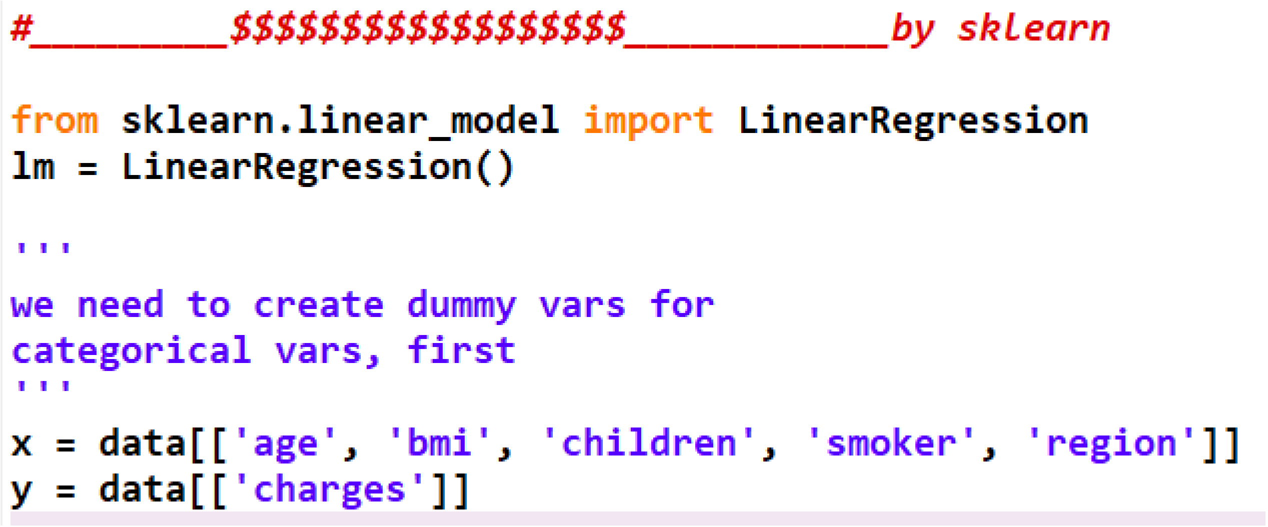
Sex should be removed

Sex removed

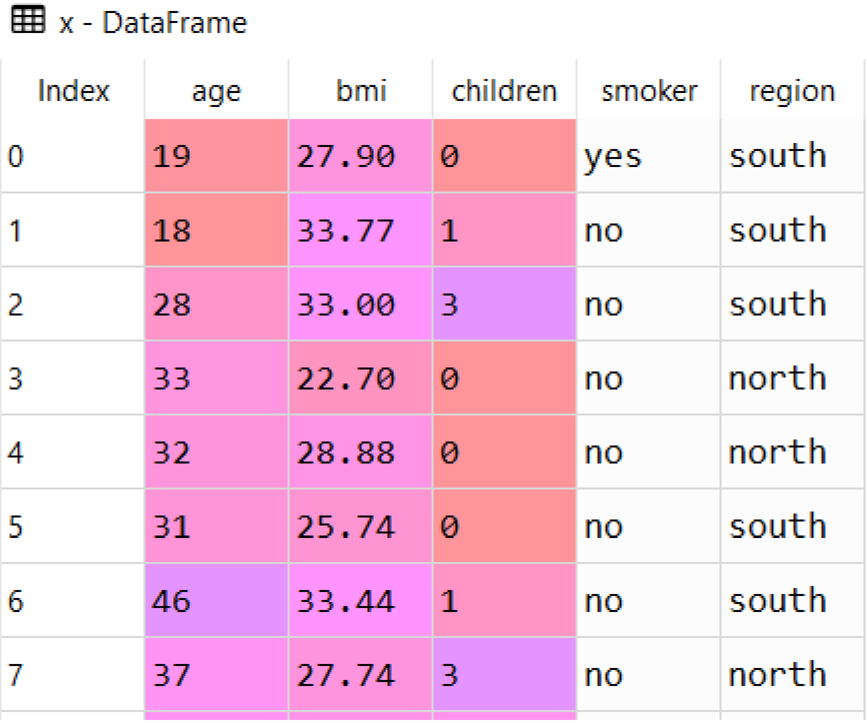


```
In [22]: model3=smf.ols(formula='charges ~ age + bmi + children + smoker + region',data=data).fit()

In [23]: print(model3.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     802.5
Date:                Mon, 27 Sep 2021   Prob (F-statistic):               0.00
Time:                        10:02:53   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1332   BIC:                         2.714e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================

==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        -1.214e+04    940.460    -12.914      0.000    -1.4e+04   -1.03e+04
smoker[T.yes]     2.384e+04    410.651     58.053      0.000     2.3e+04    2.46e+04
region[T.south]   -820.4018    341.155     -2.405      0.016   -1489.662    -151.141
age                257.0636     11.880     21.639      0.000     233.759     280.368
bmi                337.8595     28.128     12.012      0.000     282.680     393.039
children           472.1952    137.546      3.433      0.001     202.364     742.026
==============================================================================
Omnibus:                      299.848   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              715.565
Skew:                           1.210   Prob(JB):                    4.14e-156
Kurtosis:                       5.641   Cond. No.                         292.
==============================================================================

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

Sex not included, region in 2 catgs

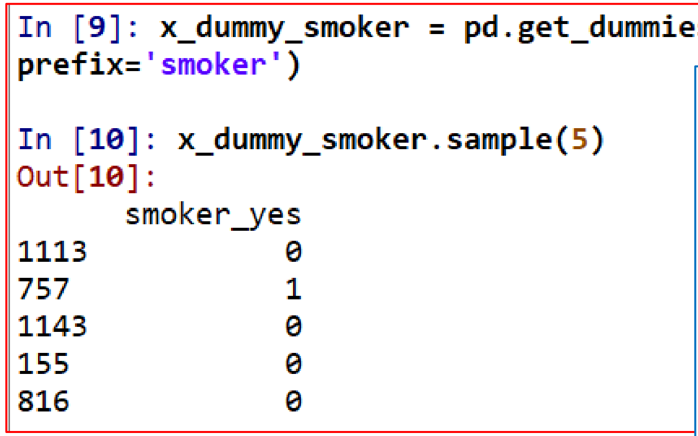


```
#_________$$$$$$$$$$$$$$$$$$____________by sklearn

from sklearn.linear_model import LinearRegression
lm = LinearRegression()

'''
we need to create dummy vars for
categorical vars, first
'''
x = data[['age', 'bmi', 'children', 'smoker', 'region']]
y = data[['charges']]
```

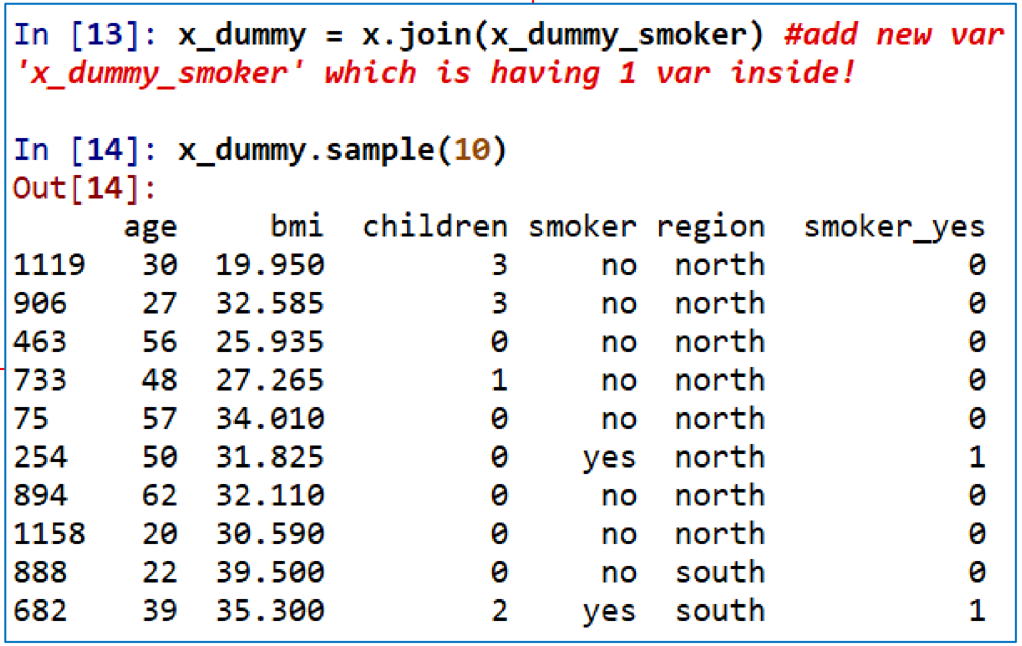


x - DataFrame

Index	age	bmi	children	smoker	region
0	19	27.90	0	yes	south
1	18	33.77	1	no	south
2	28	33.00	3	no	south
3	33	22.70	0	no	north
4	32	28.88	0	no	north
5	31	25.74	0	no	south
6	46	33.44	1	no	south
7	37	27.74	3	no	north

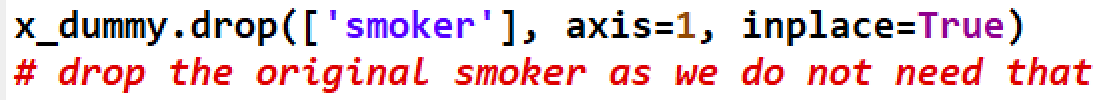


```
In [9]: x_dummy_smoker = pd.get_dummies(x.smoker, drop_first=True,
prefix='smoker')

In [10]: x_dummy_smoker.sample(5)
Out[10]:
      smoker_yes
1113           0
757            1
1143           0
155            0
816            0
```

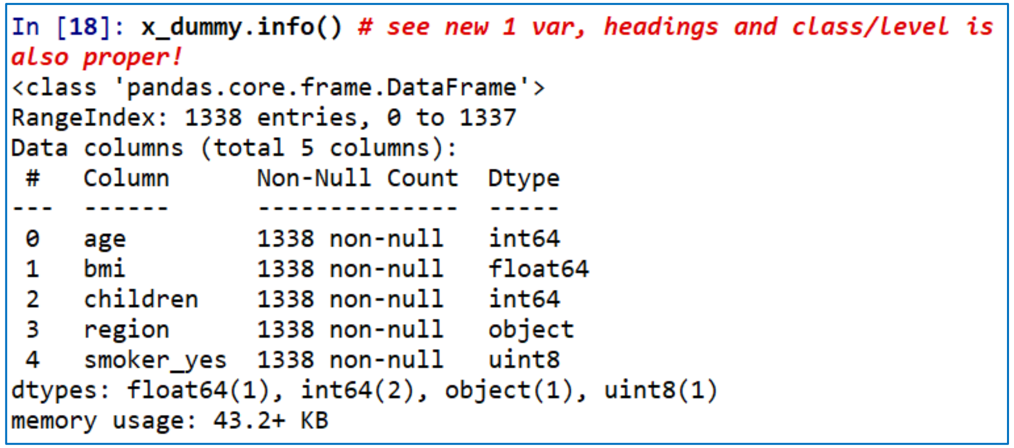


```
In [13]: x_dummy = x.join(x_dummy_smoker) #add new var
'x_dummy_smoker' which is having 1 var inside!

In [14]: x_dummy.sample(10)
Out[14]:
      age     bmi  children smoker region  smoker_yes
1119   30  19.950         3     no  north           0
906    27  32.585         3     no  north           0
463    56  25.935         0     no  north           0
733    48  27.265         1     no  north           0
75     57  34.010         0     no  north           0
254    50  31.825         0    yes  north           1
894    62  32.110         0     no  north           0
1158   20  30.590         0     no  north           0
888    22  39.500         0     no  south           0
682    39  35.300         2    yes  south           1
```

Dummy of smoker

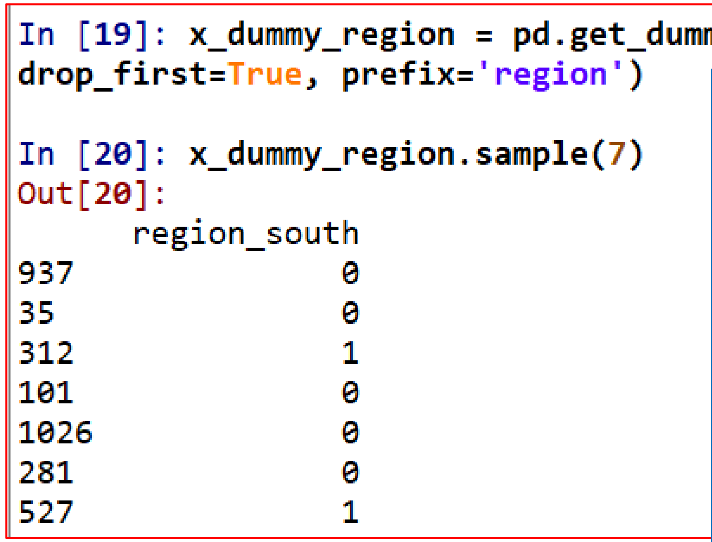


```
x_dummy.drop(['smoker'], axis=1, inplace=True)
# drop the original smoker as we do not need that
```

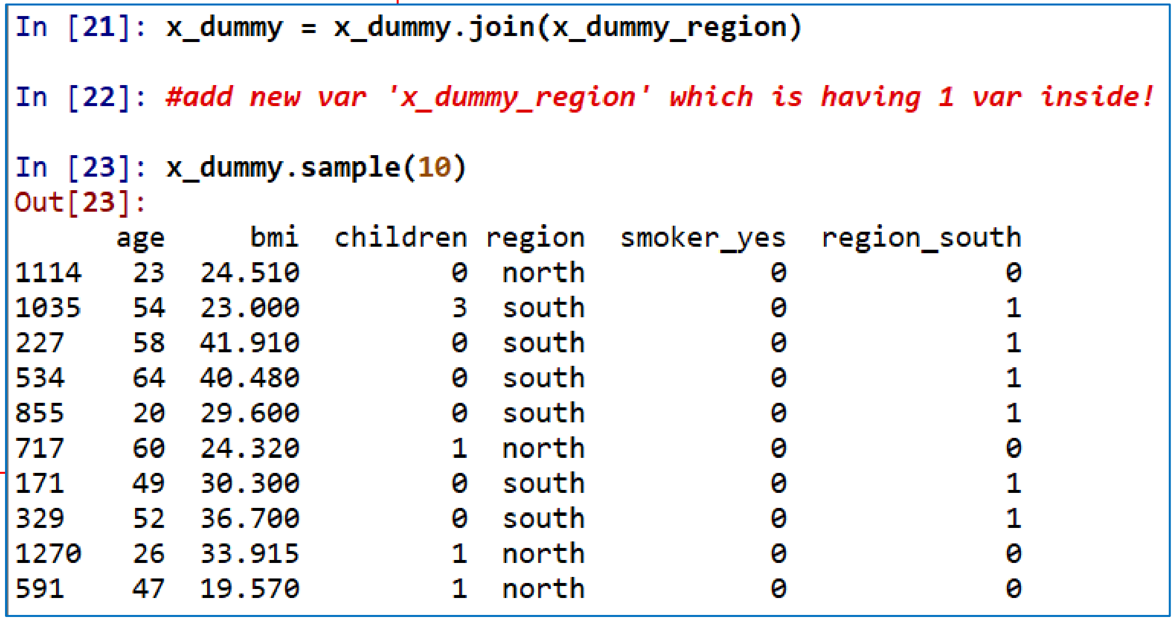


```
In [18]: x_dummy.info() # see new 1 var, headings and class/level is
also proper!
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   age         1338 non-null   int64
 1   bmi         1338 non-null   float64
 2   children    1338 non-null   int64
 3   region      1338 non-null   object
 4   smoker_yes  1338 non-null   uint8
dtypes: float64(1), int64(2), object(1), uint8(1)
memory usage: 43.2+ KB
```

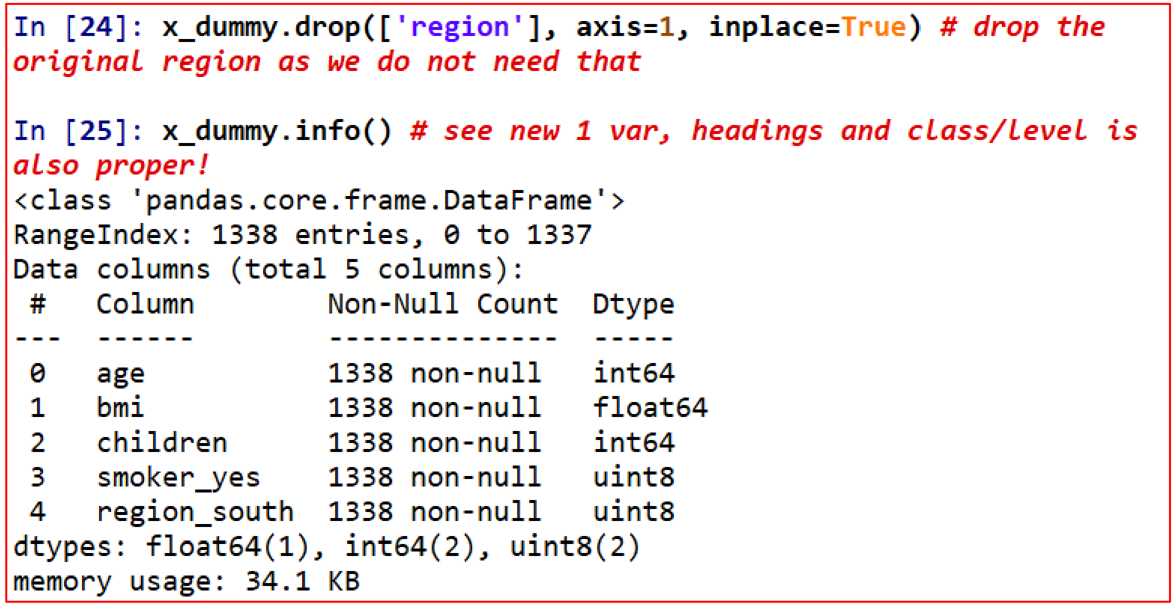
Region dummy



```
In [19]: x_dummy_region = pd.get_dummies(x_dummy.region,
drop_first=True, prefix='region')

In [20]: x_dummy_region.sample(7)
Out[20]:
      region_south
937              0
35               0
312              1
101              0
1026             0
281              0
527              1
```

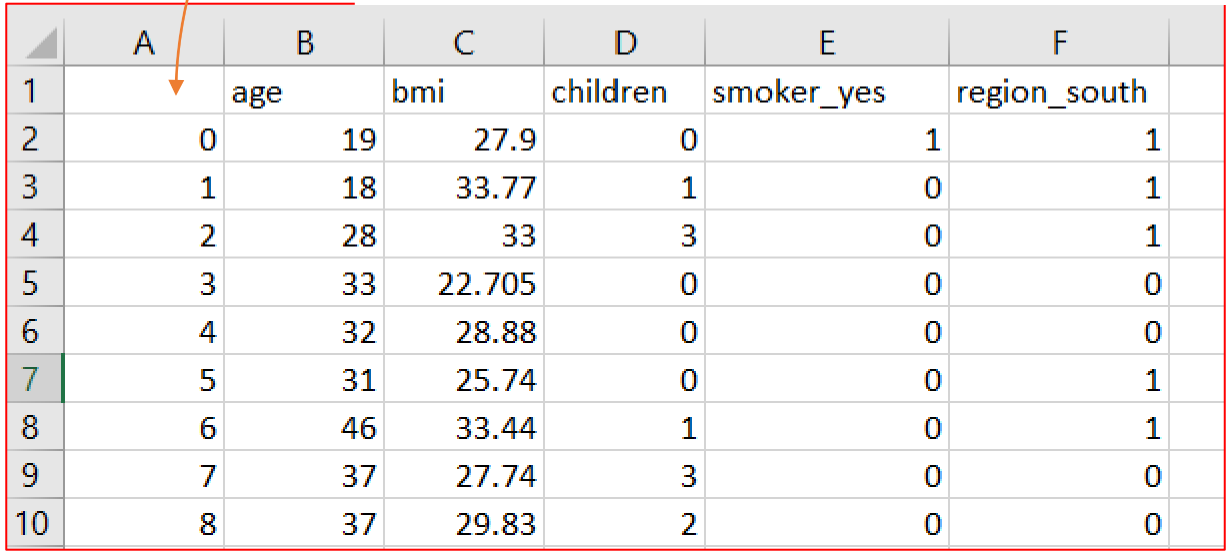


```
In [21]: x_dummy = x_dummy.join(x_dummy_region)

In [22]: #add new var 'x_dummy_region' which is having 1 var inside!

In [23]: x_dummy.sample(10)
Out[23]:
      age     bmi  children region  smoker_yes  region_south
1114   23  24.510         0  north           0             0
1035   54  23.000         3  south           0             1
227    58  41.910         0  south           0             1
534    64  40.480         0  south           0             1
855    20  29.600         0  south           0             1
717    60  24.320         1  north           0             0
171    49  30.300         0  south           0             1
329    52  36.700         0  south           0             1
1270   26  33.915         1  north           0             0
591    47  19.570         1  north           0             0
```

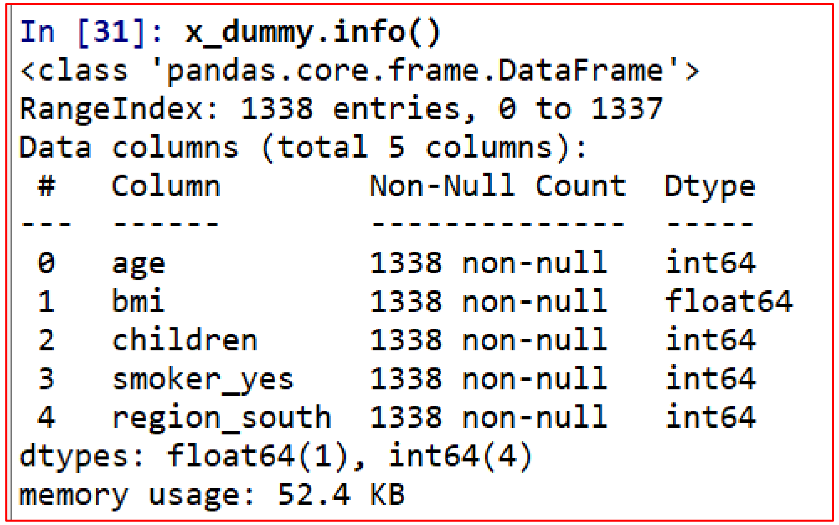


```
In [24]: x_dummy.drop(['region'], axis=1, inplace=True) # drop the
original region as we do not need that

In [25]: x_dummy.info() # see new 1 var, headings and class/level is
also proper!
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 5 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   smoker_yes    1338 non-null   uint8
 4   region_south  1338 non-null   uint8
dtypes: float64(1), int64(2), uint8(2)
memory usage: 34.1 KB
```

```
# as we dont want to repeat the previous codes again, in case
# let us save at desktop/folder and reimport

x_dummy.to_csv("C:/Users/Dr Vinod/Desktop/x_dummy.csv")
# go to file and delete 1st column having serial nos

x_dummy = pd.read_csv("C:/Users/Dr Vinod/Desktop/x_dummy.csv")
x_dummy.info()
```



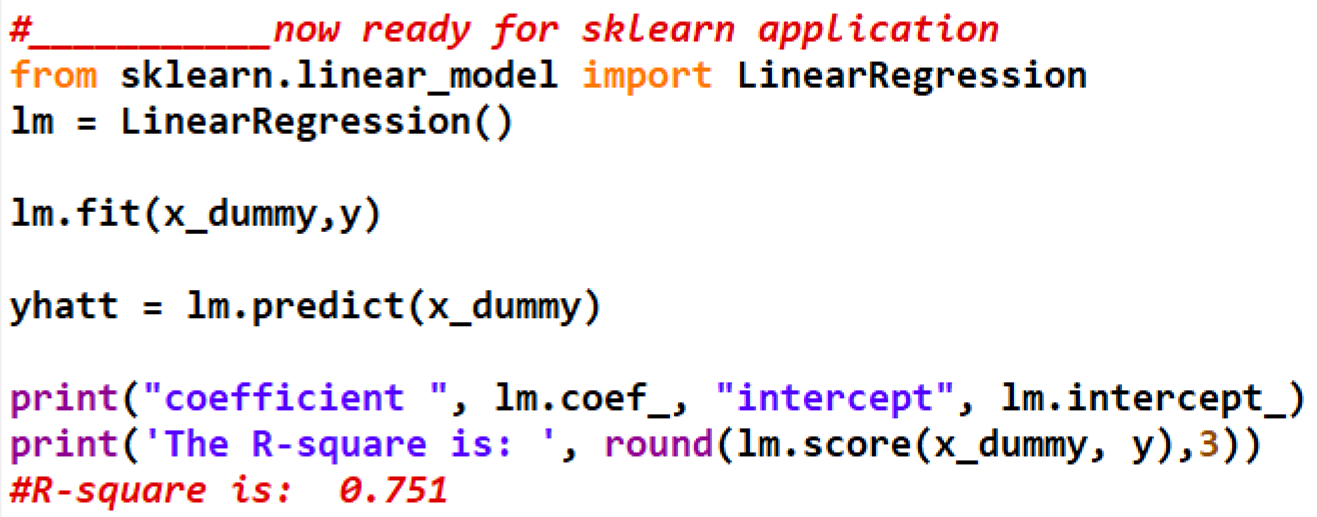
	A	B	C	D	E	F
1		age	bmi	children	smoker_yes	region_south
2	0	19	27.9	0	1	1
3	1	18	33.77	1	0	1
4	2	28	33	3	0	1
5	3	33	22.705	0	0	0
6	4	32	28.88	0	0	0
7	5	31	25.74	0	0	1
8	6	46	33.44	1	0	1
9	7	37	27.74	3	0	0
10	8	37	29.83	2	0	0



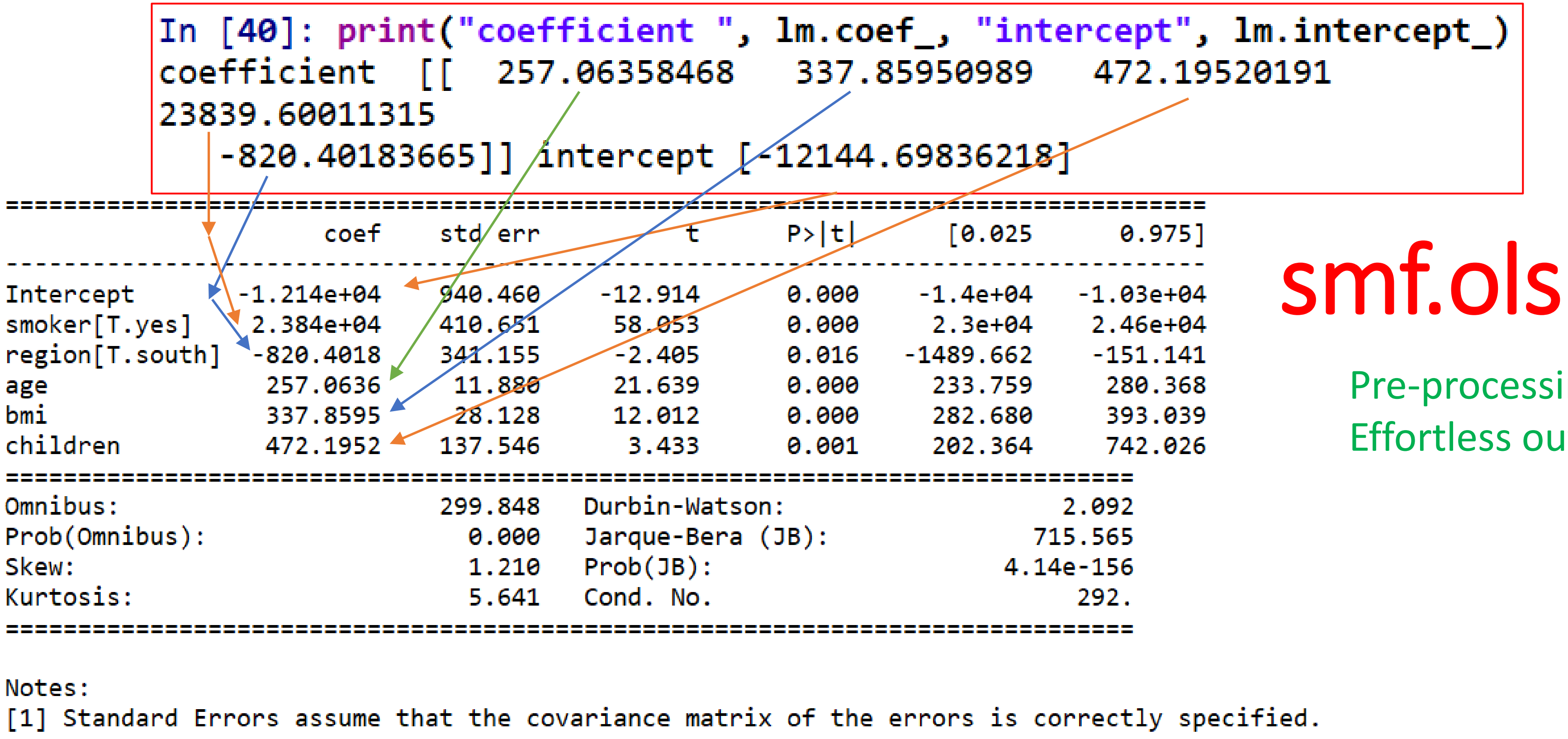
```
In [31]: x_dummy.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   age           1338 non-null   int64  
 1   bmi           1338 non-null   float64
 2   children      1338 non-null   int64  
 3   smoker_yes    1338 non-null   int64  
 4   region_south  1338 non-null   int64  
dtypes: float64(1), int64(4)
memory usage: 52.4 KB
```

Sklearn application

Pre-processing required
Efforts required for
output



```
#___________now ready for sklearn application
from sklearn.linear_model import LinearRegression
lm = LinearRegression()

lm.fit(x_dummy,y)

yhatt = lm.predict(x_dummy)

print("coefficient ", lm.coef_, "intercept", lm.intercept_)
print('The R-square is: ', round(lm.score(x_dummy, y),3))
#R-square is:  0.751
```



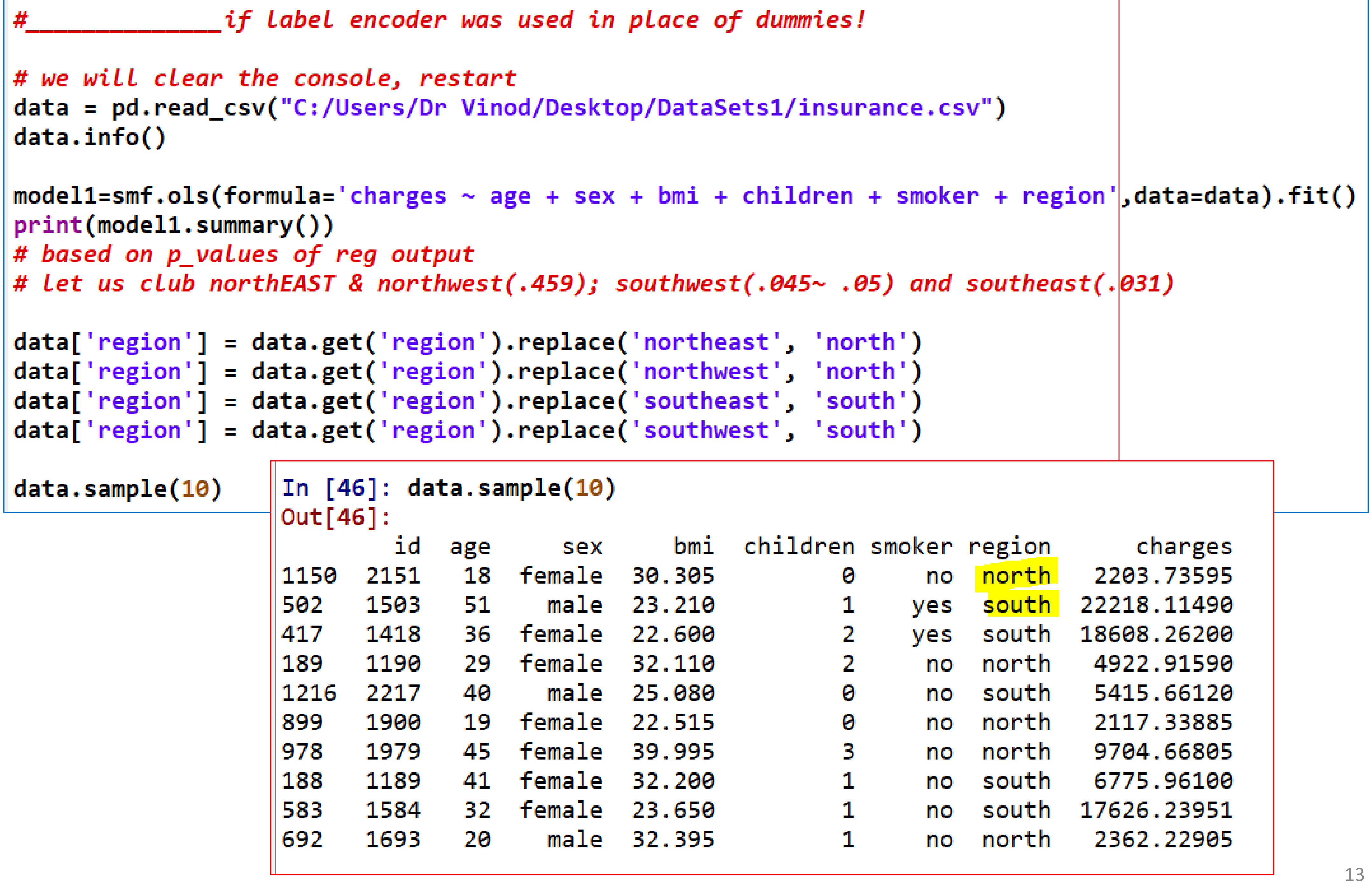
```
In [40]: print("coefficient ", lm.coef_, "intercept", lm.intercept_)
coefficient  [[  257.06358468   337.85950989   472.19520191
23839.60011315
   -820.40183665]] intercept [-12144.69836218]
```

```
======================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept      -1.214e+04    940.460    -12.914      0.000    -1.4e+04   -1.03e+04
smoker[T.yes]   2.384e+04    410.651     58.053      0.000     2.3e+04    2.46e+04
region[T.south] -820.4018    341.155     -2.405      0.016   -1489.662    -151.141
age              257.0636     11.880     21.639      0.000     233.759     280.368
bmi              337.8595     28.128     12.012      0.000     282.680     393.039
children         472.1952    137.546      3.433      0.001     202.364     742.026
==============================================================================
Omnibus:                      299.848   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              715.565
Skew:                           1.210   Prob(JB):                    4.14e-156
Kurtosis:                       5.641   Cond. No.                         292.
==============================================================================

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

smf.ols application

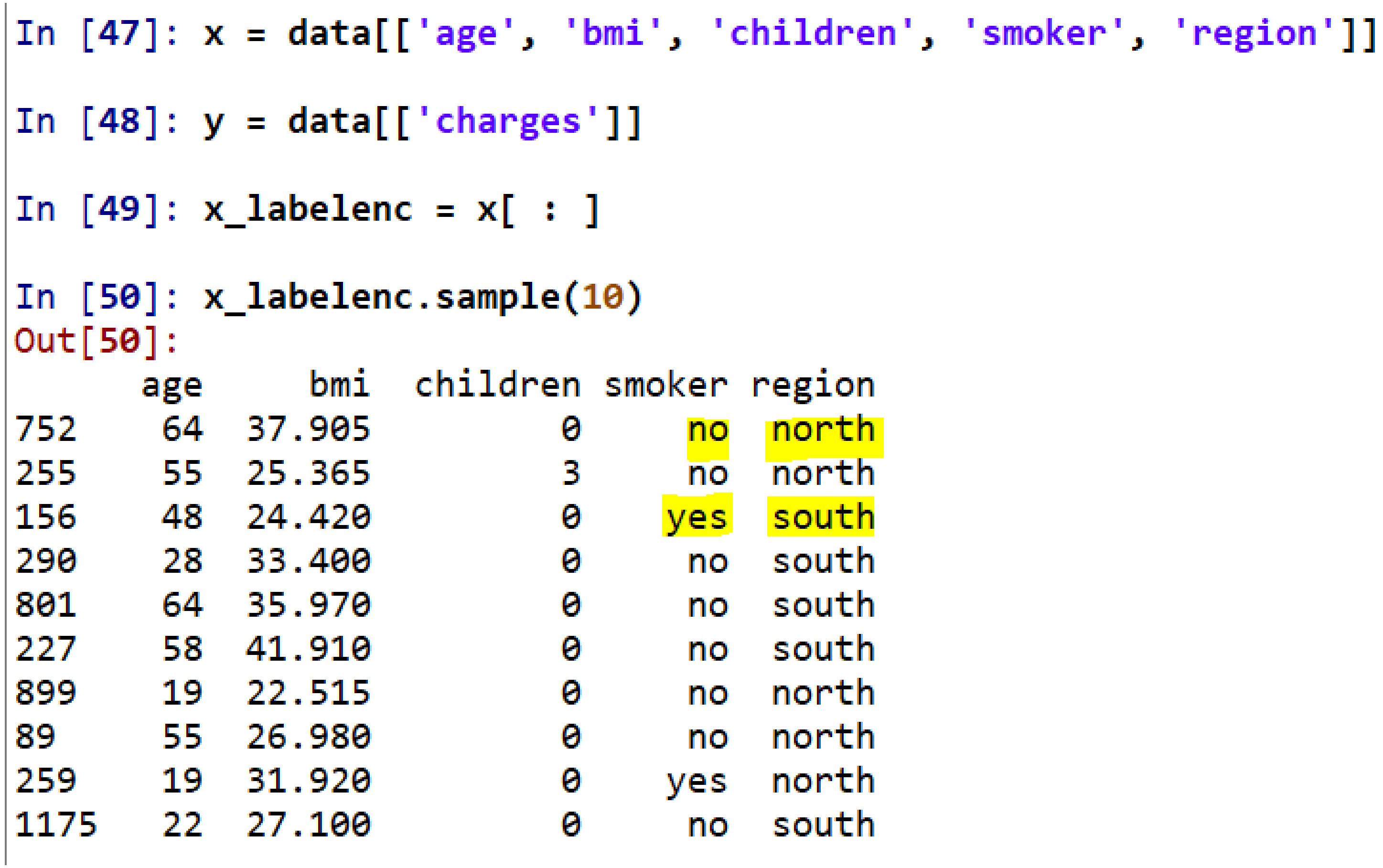
Pre-processing NOT required
Effortless output

What if, we use label encoding!

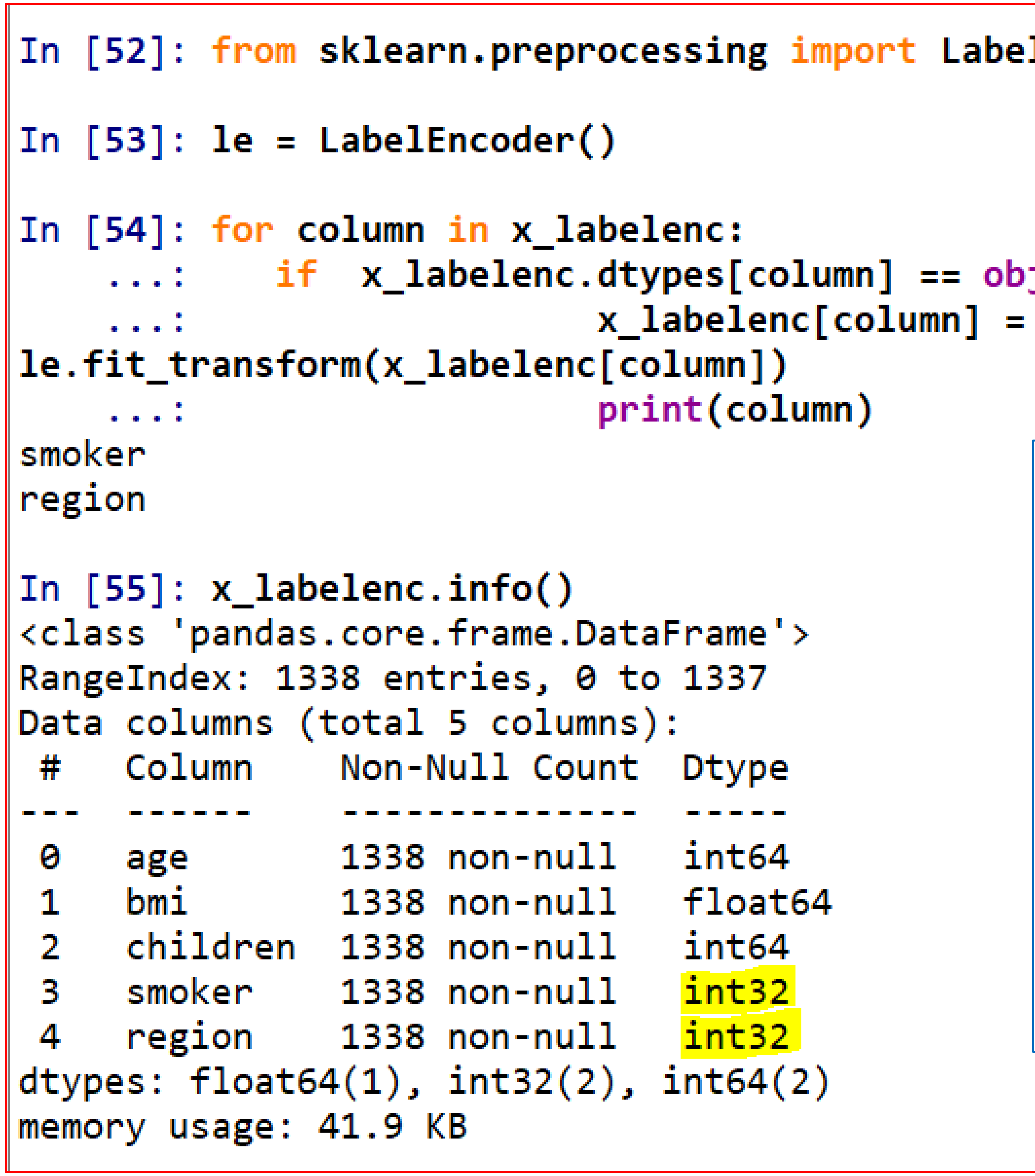


```
#______________if label encoder was used in place of dummies!

# we will clear the console, restart
data = pd.read_csv("C:/Users/Dr Vinod/Desktop/DataSets1/insurance.csv")
data.info()

model1=smf.ols(formula='charges ~ age + sex + bmi + children + smoker + region',data=data).fit()
print(model1.summary())
# based on p_values of reg output
# let us club northEAST & northwest(.459); southwest(.045~ .05) and southeast(.031)

data['region'] = data.get('region').replace('northeast', 'north')
data['region'] = data.get('region').replace('northwest', 'north')
data['region'] = data.get('region').replace('southeast', 'south')
data['region'] = data.get('region').replace('southwest', 'south')

data.sample(10)
```

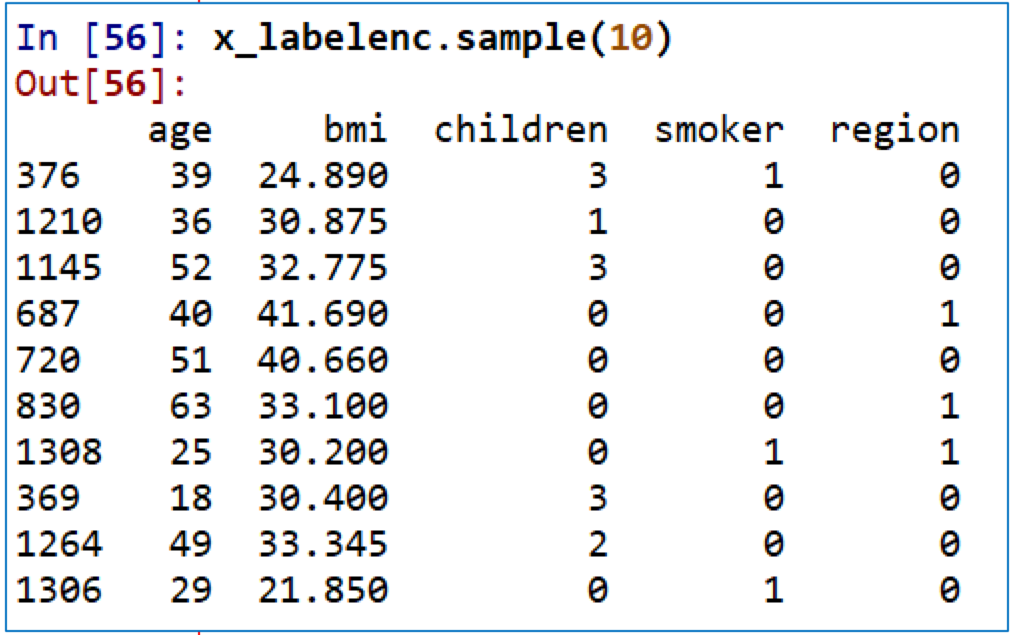
```
In [46]: data.sample(10)
Out[46]:
        id  age     sex     bmi  children smoker region      charges
1150  2151   18  female  30.305         0     no  north   2203.73595
502   1503   51    male  23.210         1    yes  south  22218.11490
417   1418   36  female  22.600         2    yes  south  18608.26200
189   1190   29  female  32.110         2     no  north   4922.91590
1216  2217   40    male  25.080         0     no  south   5415.66120
899   1900   19  female  22.515         0     no  north   2117.33885
978   1979   45  female  39.995         3     no  north   9704.66805
188   1189   41  female  32.200         1     no  south   6775.96100
583   1584   32  female  23.650         1     no  south  17626.23951
692   1693   20    male  32.395         1     no  north   2362.22905
```



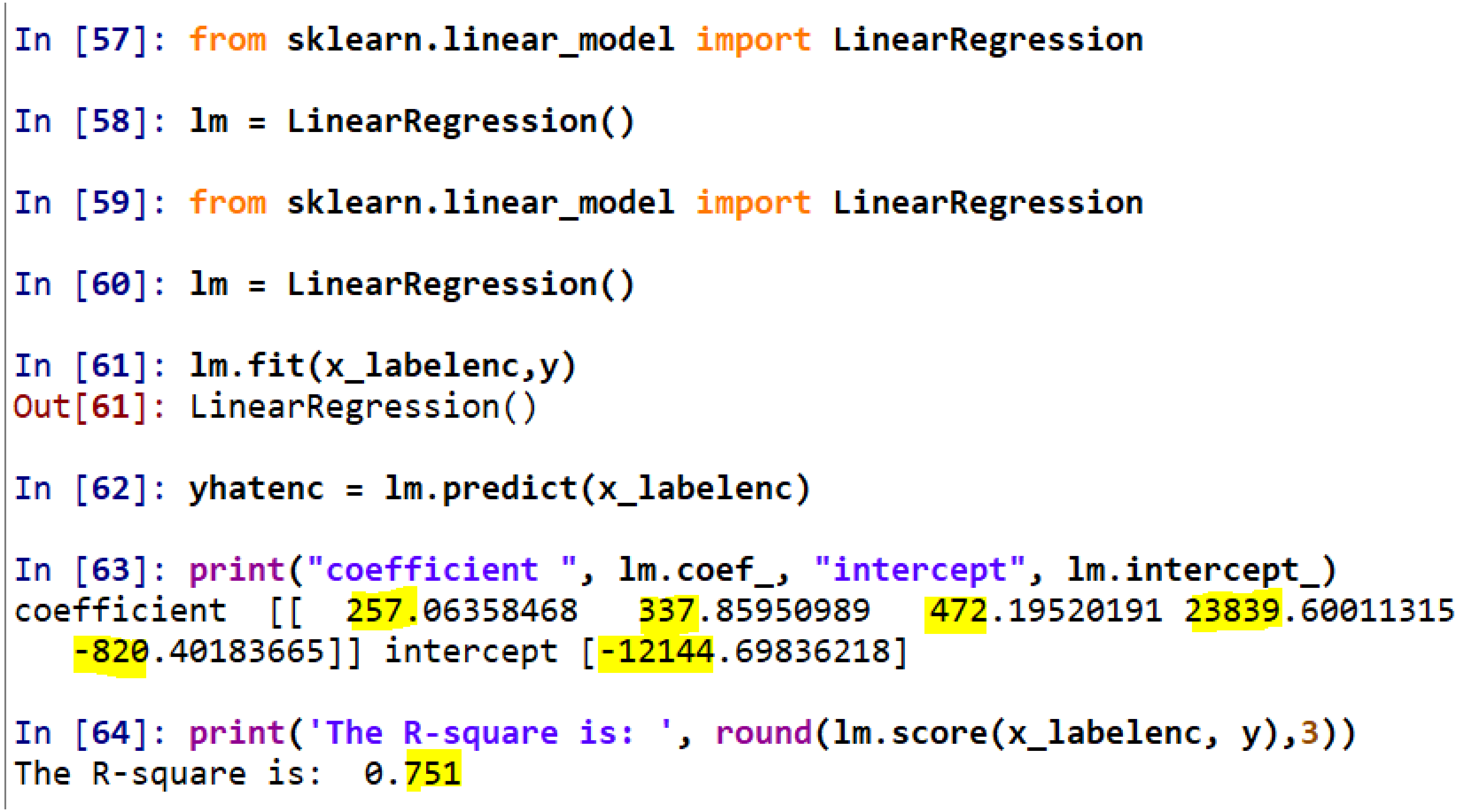
```
In [47]: x = data[['age', 'bmi', 'children', 'smoker', 'region']]

In [48]: y = data[['charges']]

In [49]: x_labelenc = x[ : ]

In [50]: x_labelenc.sample(10)
Out[50]: 
      age     bmi  children smoker region
752    64  37.905         0     no  north
255    55  25.365         3     no  north
156    48  24.420         0    yes  south
290    28  33.400         0     no  south
801    64  35.970         0     no  south
227    58  41.910         0     no  south
899    19  22.515         0     no  north
89     55  26.980         0     no  north
259    19  31.920         0    yes  north
1175   22  27.100         0     no  south
```



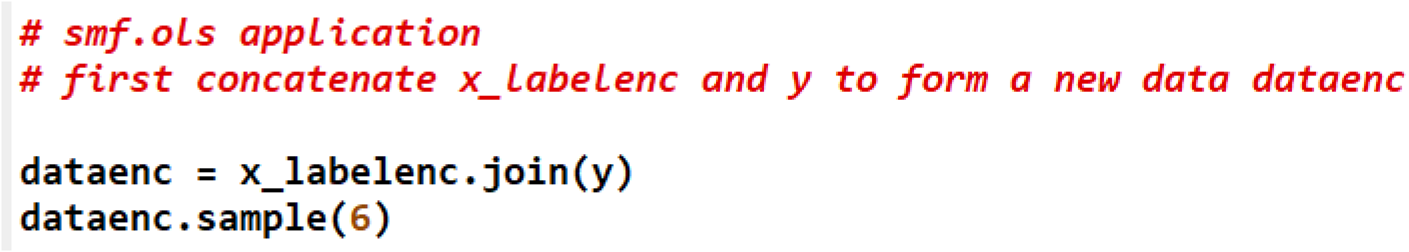
```
In [52]: from sklearn.preprocessing import LabelEncoder

In [53]: le = LabelEncoder()

In [54]: for column in x_labelenc:
    ...:     if  x_labelenc.dtypes[column] == object:
    ...:                     x_labelenc[column] =
le.fit_transform(x_labelenc[column])
    ...:                     print(column)
smoker
region

In [55]: x_labelenc.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 5 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   smoker    1338 non-null   int32  
 4   region    1338 non-null   int32  
dtypes: float64(1), int32(2), int64(2)
memory usage: 41.9 KB
```



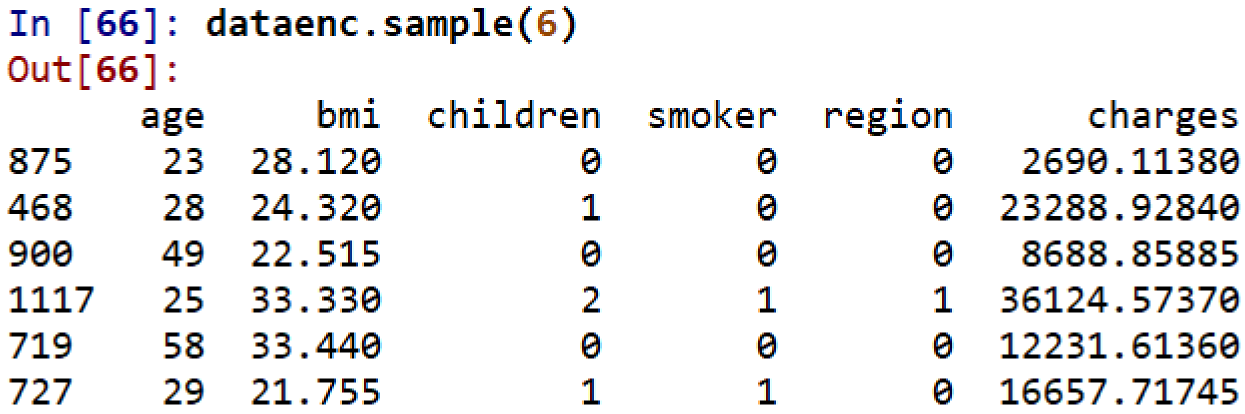
```
In [56]: x_labelenc.sample(10)
Out[56]: 
      age     bmi  children  smoker  region
376    39  24.890         3       1       0
1210   36  30.875         1       0       0
1145   52  32.775         3       0       0
687    40  41.690         0       0       1
720    51  40.660         0       0       0
830    63  33.100         0       0       1
1308   25  30.200         0       1       1
369    18  30.400         3       0       0
1264   49  33.345         2       0       0
1306   29  21.850         0       1       0
```



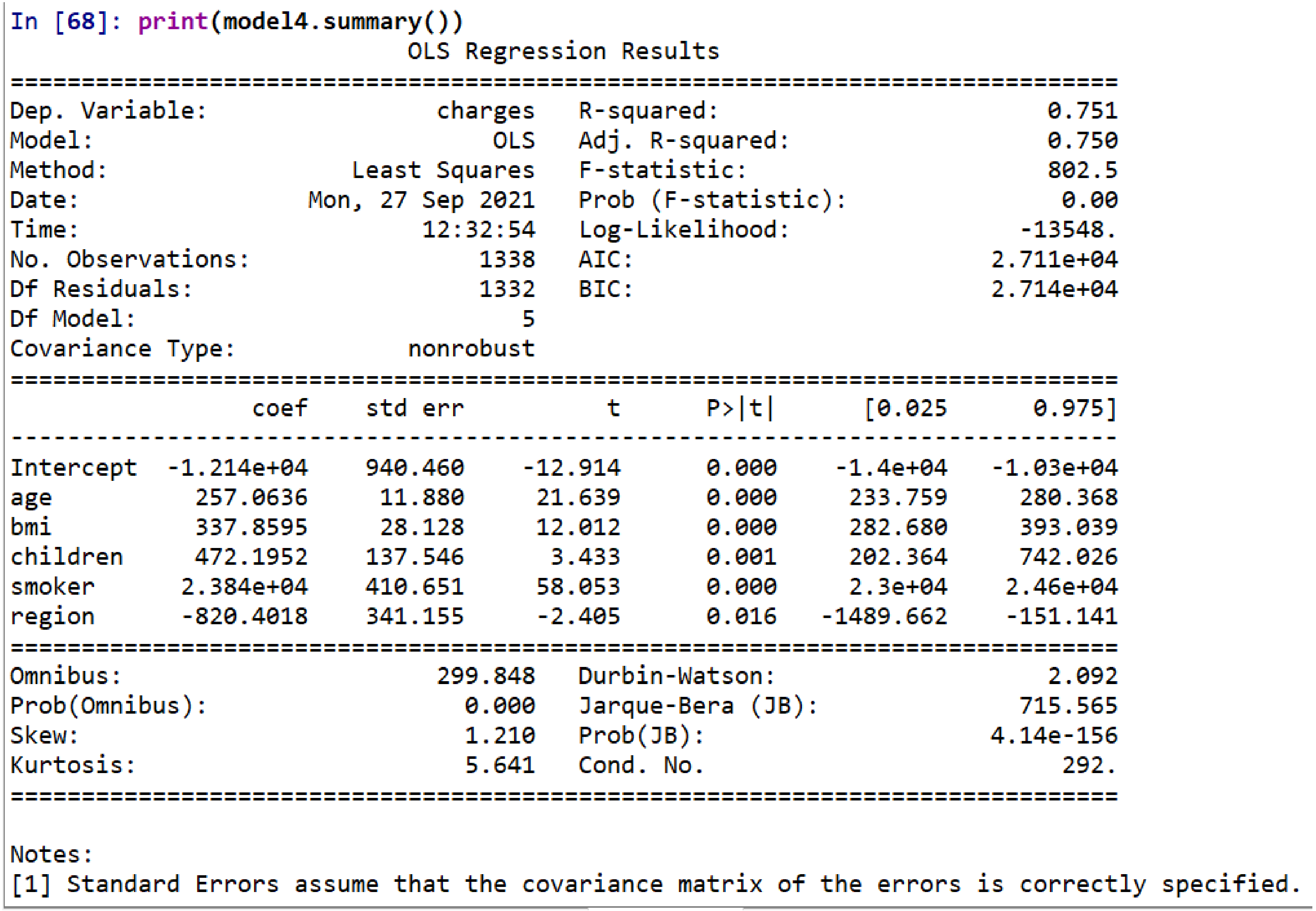
```
In [57]: from sklearn.linear_model import LinearRegression

In [58]: lm = LinearRegression()

In [59]: from sklearn.linear_model import LinearRegression

In [60]: lm = LinearRegression()

In [61]: lm.fit(x_labelenc,y)
Out[61]: LinearRegression()

In [62]: yhatenc = lm.predict(x_labelenc)

In [63]: print("coefficient ", lm.coef_, "intercept", lm.intercept_)
coefficient  [[  257.06358468   337.85950989   472.19520191 23839.60011315
   -820.40183665]] intercept [-12144.69836218]

In [64]: print('The R-square is: ', round(lm.score(x_labelenc, y),3))
The R-square is:  0.751
```



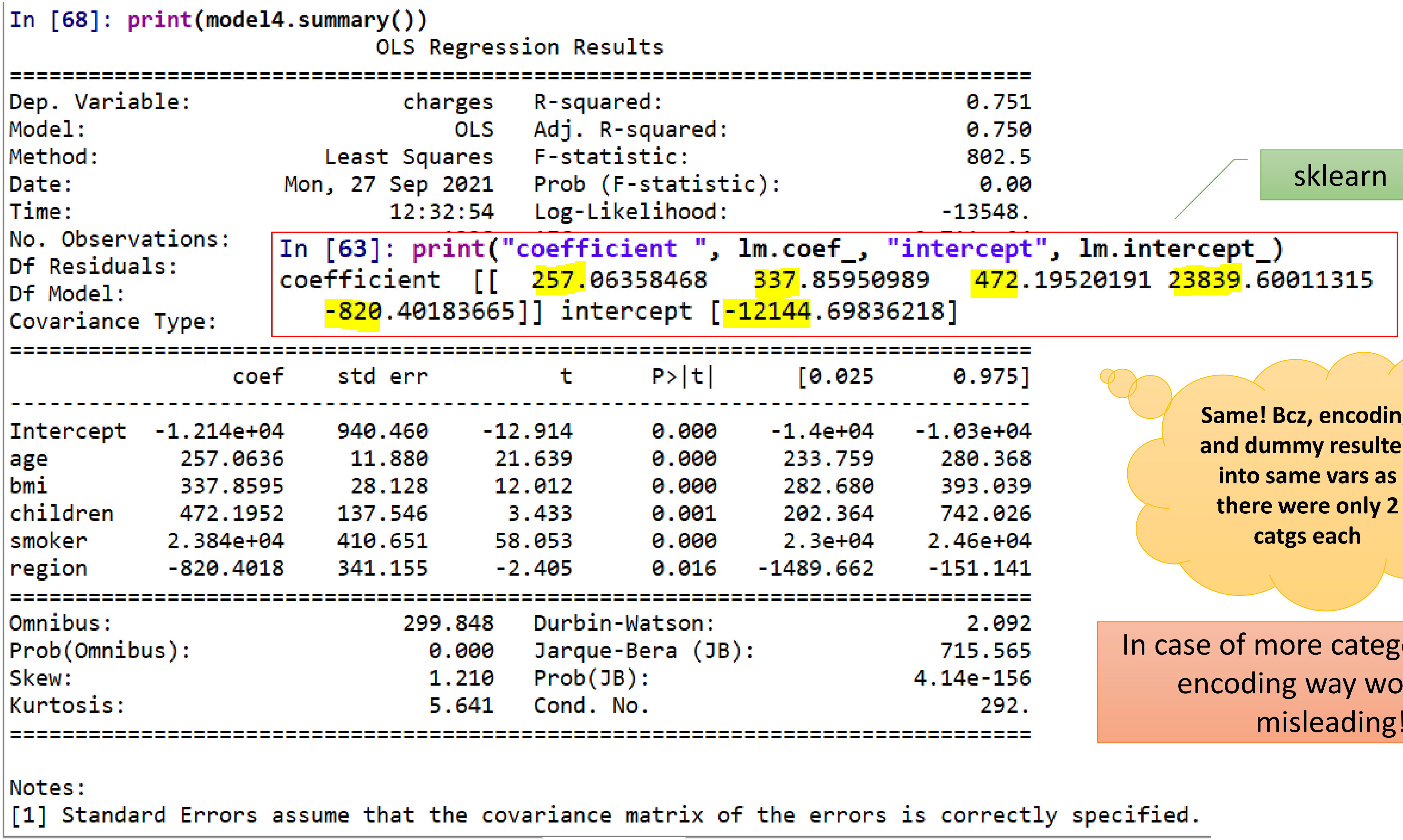
```
# smf.ols application
# first concatenate x_labelenc and y to form a new data dataenc

dataenc = x_labelenc.join(y)
dataenc.sample(6)
```



```
In [66]: dataenc.sample(6)
Out[66]:
      age     bmi  children  smoker  region      charges
875    23  28.120         0       0       0   2690.11380
468    28  24.320         1       0       0  23288.92840
900    49  22.515         0       0       0   8688.85885
1117   25  33.330         2       1       1  36124.57370
719    58  33.440         0       0       0  12231.61360
727    29  21.755         1       1       0  16657.71745
```



```
In [68]: print(model4.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     802.5
Date:                Mon, 27 Sep 2021   Prob (F-statistic):               0.00
Time:                        12:32:54   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1332   BIC:                         2.714e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.214e+04    940.460    -12.914      0.000    -1.4e+04   -1.03e+04
age          257.0636     11.880     21.639      0.000     233.759     280.368
bmi          337.8595     28.128     12.012      0.000     282.680     393.039
children     472.1952    137.546      3.433      0.001     202.364     742.026
smoker      2.384e+04    410.651     58.053      0.000     2.3e+04    2.46e+04
region      -820.4018    341.155     -2.405      0.016   -1489.662    -151.141
==============================================================================
Omnibus:                      299.848   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              715.565
Skew:                           1.210   Prob(JB):                    4.14e-156
Kurtosis:                       5.641   Cond. No.                         292.
==============================================================================

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



```
In [68]: print(model4.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     802.5
Date:                Mon, 27 Sep 2021   Prob (F-statistic):               0.00
Time:                        12:32:54   Log-Likelihood:                -13548.
No. Observations:
Df Residuals:
Df Model:
Covariance Type:
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.214e+04    940.460    -12.914      0.000    -1.4e+04   -1.03e+04
age          257.0636     11.880     21.639      0.000     233.759     280.368
bmi          337.8595     28.128     12.012      0.000     282.680     393.039
children     472.1952    137.546      3.433      0.001     202.364     742.026
smoker      2.384e+04    410.651     58.053      0.000      2.3e+04    2.46e+04
region      -820.4018    341.155     -2.405      0.016   -1489.662    -151.141
==============================================================================
Omnibus:                      299.848   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              715.565
Skew:                           1.210   Prob(JB):                    4.14e-156
Kurtosis:                       5.641   Cond. No.                         292.
==============================================================================

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

```
In [63]: print("coefficient ", lm.coef_, "intercept", lm.intercept_)
coefficient  [[  257.06358468   337.85950989   472.19520191 23839.60011315
   -820.40183665]] intercept [-12144.69836218]
```

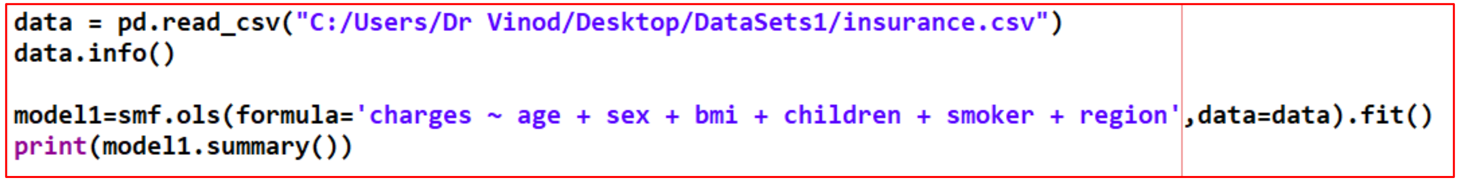
sklearn

Same! Bcz, encoding and dummy resulted into same vars as there were only 2 catgs each

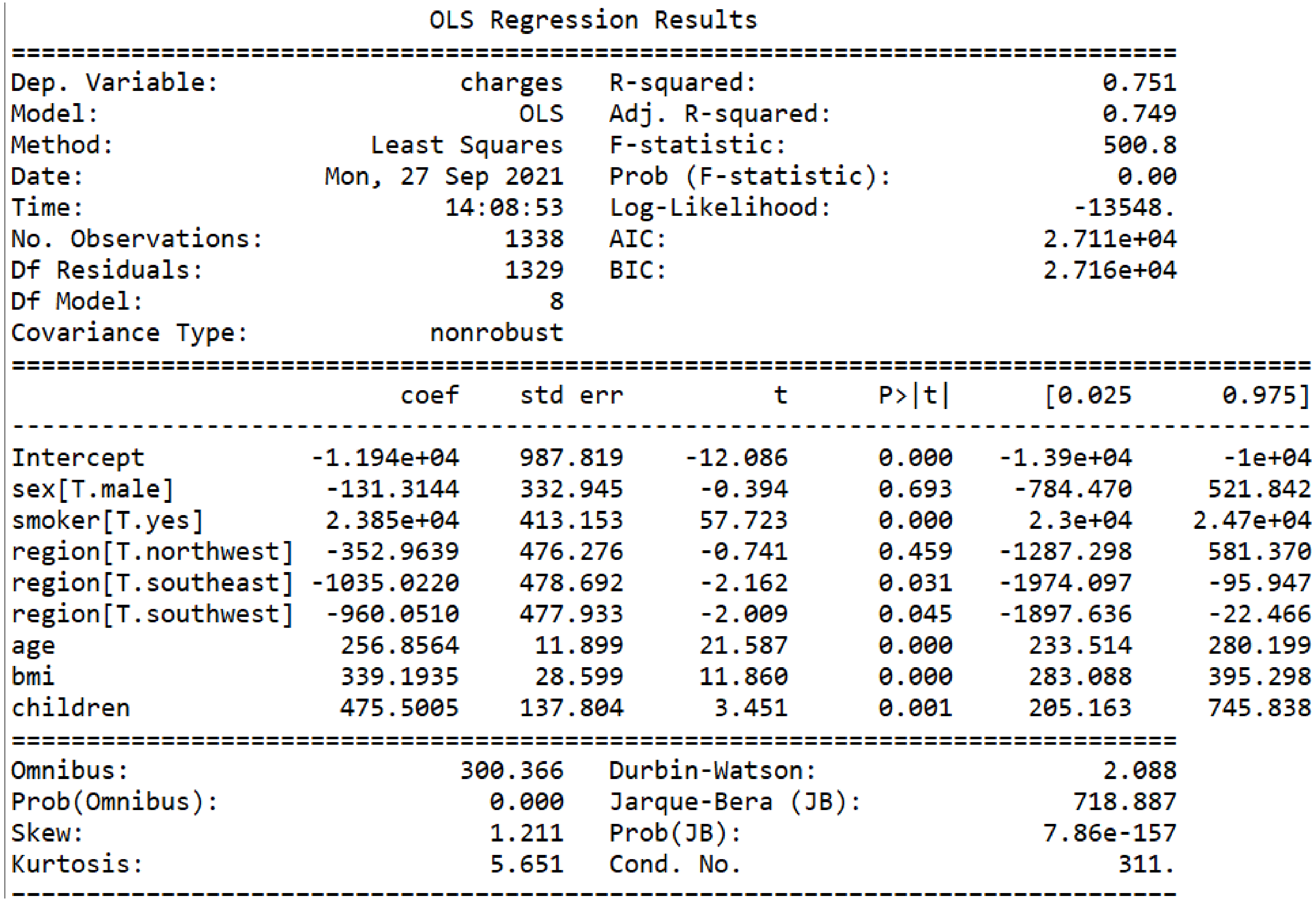
In case of more categories, then encoding way would be misleading!

Let’s do label encoding when levels in categorical are more than 2

This was our 1st model

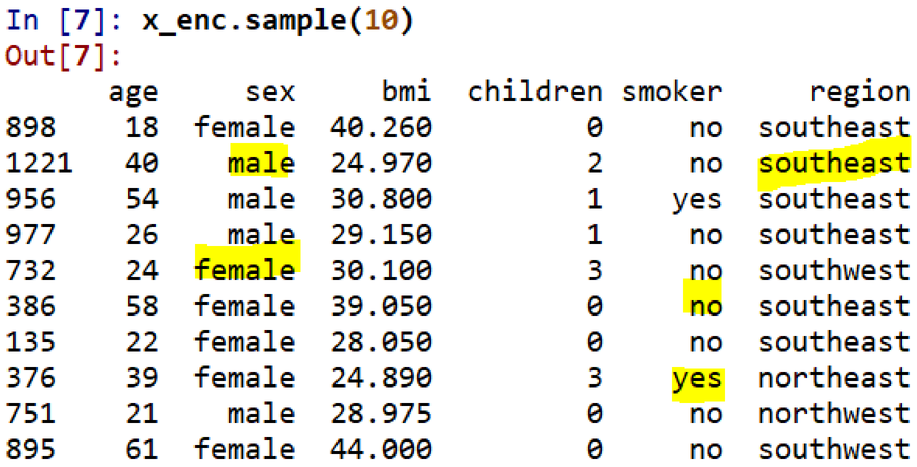


```
data = pd.read_csv("C:/Users/Dr Vinod/Desktop/DataSets1/insurance.csv")
data.info()

model1=smf.ols(formula='charges ~ age + sex + bmi + children + smoker + region',data=data).fit()
print(model1.summary())
```



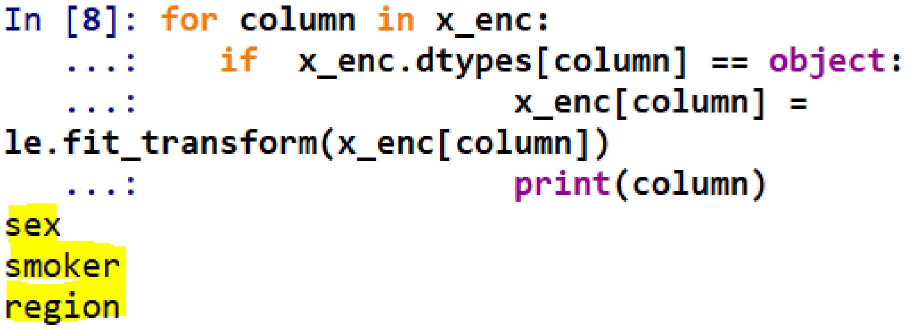
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     500.8
Date:                Mon, 27 Sep 2021   Prob (F-statistic):               0.00
Time:                        14:08:53   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1329   BIC:                         2.716e+04
Df Model:                           8
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept           -1.194e+04    987.819    -12.086      0.000   -1.39e+04      -1e+04
sex[T.male]          -131.3144    332.945     -0.394      0.693    -784.470     521.842
smoker[T.yes]        2.385e+04    413.153     57.723      0.000     2.3e+04    2.47e+04
region[T.northwest]  -352.9639    476.276     -0.741      0.459   -1287.298     581.370
region[T.southeast] -1035.0220    478.692     -2.162      0.031   -1974.097     -95.947
region[T.southwest]  -960.0510    477.933     -2.009      0.045   -1897.636     -22.466
age                   256.8564     11.899     21.587      0.000     233.514     280.199
bmi                   339.1935     28.599     11.860      0.000     283.088     395.298
children              475.5005    137.804      3.451      0.001     205.163     745.838
==============================================================================
Omnibus:                      300.366   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.887
Skew:                           1.211   Prob(JB):                    7.86e-157
Kurtosis:                       5.651   Cond. No.                         311.
------------------------------------------------------------------------------
```

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

x_enc = x[ : ]

x_enc.sample(10)
```



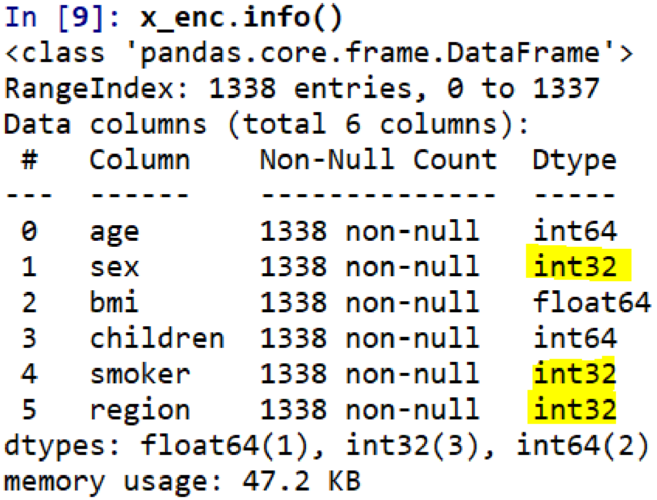
```
In [7]: x_enc.sample(10)
Out[7]:
      age     sex     bmi  children smoker     region
898    18  female  40.260         0     no  southeast
1221   40    male  24.970         2     no  southeast
956    54    male  30.800         1    yes  southeast
977    26    male  29.150         1     no  southeast
732    24  female  30.100         3     no  southwest
386    58  female  39.050         0     no  southeast
135    22  female  28.050         0     no  southeast
376    39  female  24.890         3    yes  northeast
751    21    male  28.975         0     no  northwest
895    61  female  44.000         0     no  southwest
```

```
x_enc.sample(10)

for column in x_enc:
   if  x_enc.dtypes[column] == object:
                  x_enc[column] = le.fit_transform(x_enc[column])
                  print(column)
```

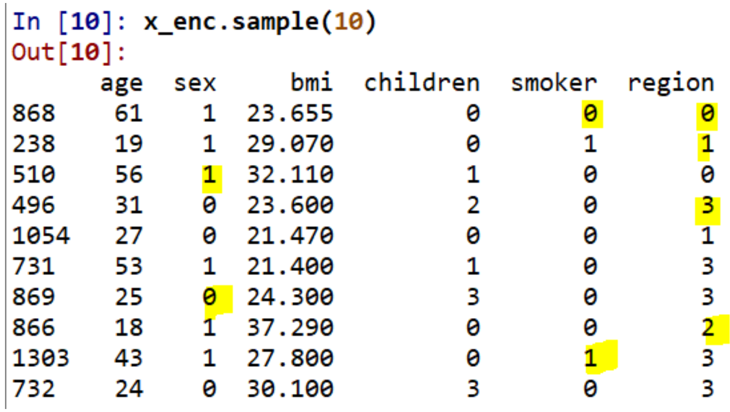


```
In [8]: for column in x_enc:
   ...:     if  x_enc.dtypes[column] == object:
   ...:                   x_enc[column] =
le.fit_transform(x_enc[column])
   ...:                   print(column)
sex
smoker
region
```

```
x_enc.info()
```



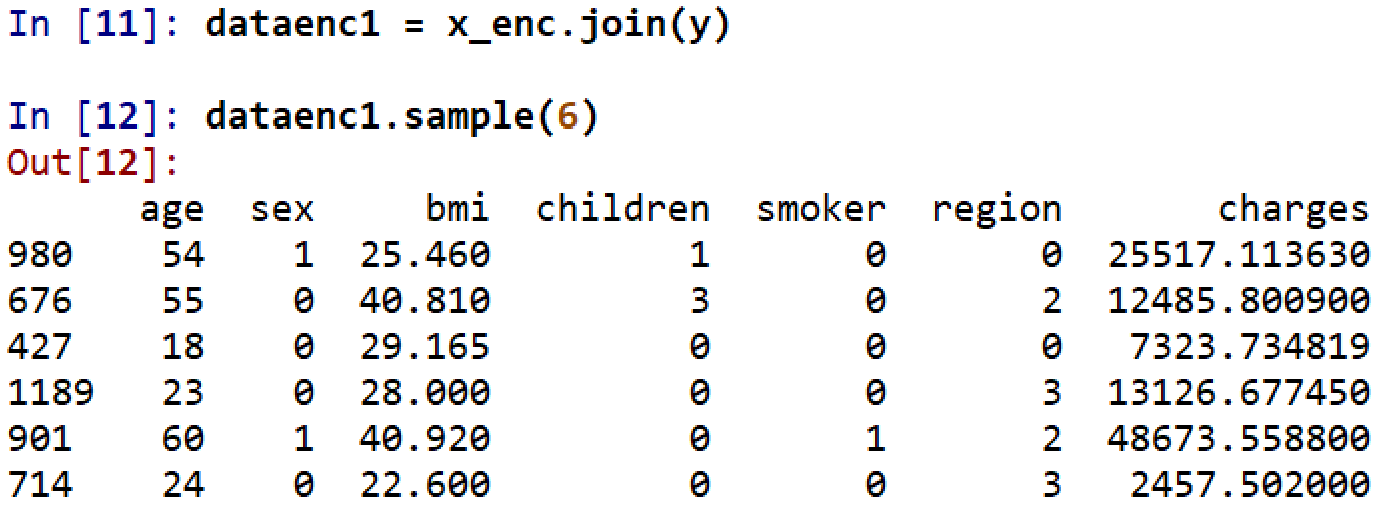
```
In [9]: x_enc.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   int32
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   int32
 5   region    1338 non-null   int32
dtypes: float64(1), int32(3), int64(2)
memory usage: 47.2 KB
```



```
In [10]: x_enc.sample(10)
Out[10]:
      age  sex     bmi  children  smoker  region
868    61    1  23.655         0       0       0
238    19    1  29.070         0       1       1
510    56    1  32.110         1       0       0
496    31    0  23.600         2       0       3
1054   27    0  21.470         0       0       1
731    53    1  21.400         1       0       3
869    25    0  24.300         3       0       3
866    18    1  37.290         0       0       2
1303   43    1  27.800         0       1       3
732    24    0  30.100         3       0       3
```

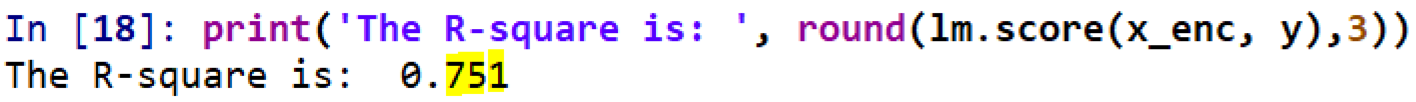
Label encoder

Join

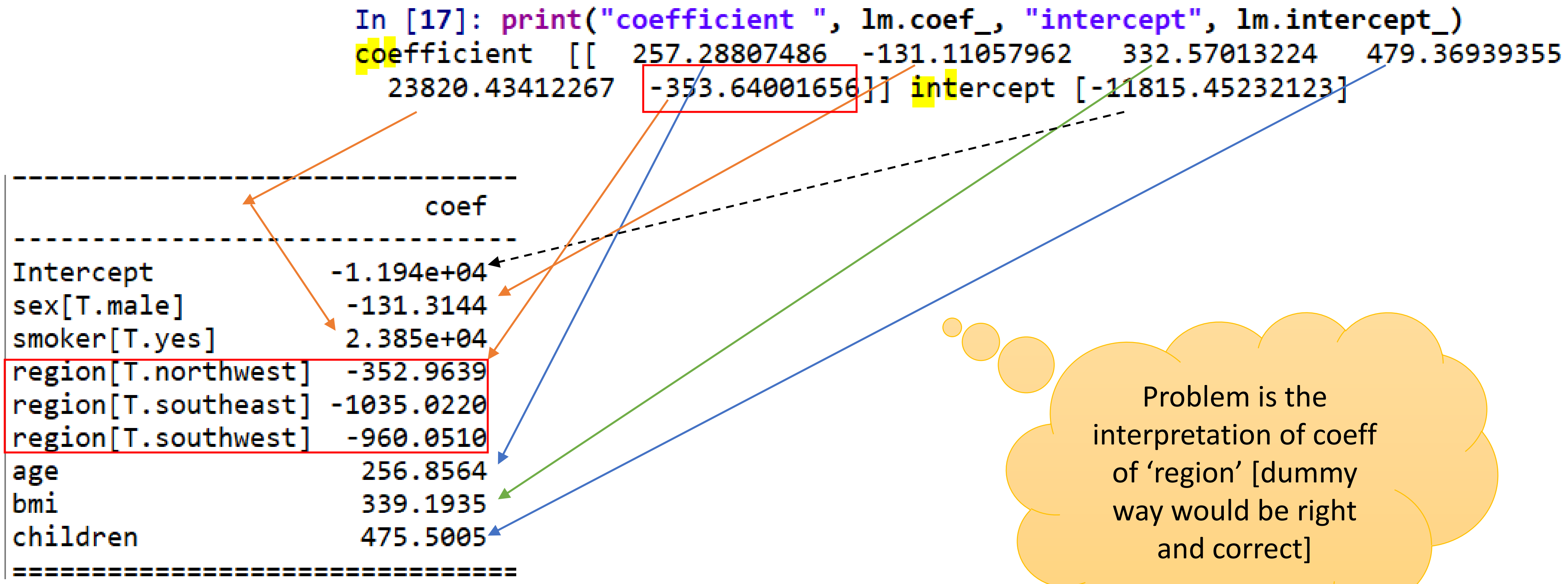


```
In [11]: dataenc1 = x_enc.join(y)

In [12]: dataenc1.sample(6)
Out[12]:
      age  sex     bmi  children  smoker  region       charges
980    54    1  25.460         1       0       0  25517.113630
676    55    0  40.810         3       0       2  12485.800900
427    18    0  29.165         0       0       0   7323.734819
1189   23    0  28.000         0       0       3  13126.677450
901    60    1  40.920         0       1       2  48673.558800
714    24    0  22.600         0       0       3   2457.502000
```

Comparison



```
In [18]: print('The R-square is: ', round(lm.score(x_enc, y),3))
The R-square is:  0.751
```



```
In [17]: print("coefficient ", lm.coef_, "intercept", lm.intercept_)
coefficient  [[  257.28807486  -131.11057962   332.57013224   479.36939355
  23820.43412267  -353.64001656]] intercept [-11815.45232123]
```

```
----------------------------------
                              coef
----------------------------------
Intercept               -1.194e+04
sex[T.male]              -131.3144
smoker[T.yes]            2.385e+04
region[T.northwest]      -352.9639
region[T.southeast]     -1035.0220
region[T.southwest]      -960.0510
age                       256.8564
bmi                       339.1935
children                  475.5005
==================================
```

Problem is the interpretation of coeff of ‘region’ [dummy way would be right and correct]