PythonRegression

November 18, 2020

1 Linear Regression in Python

[2]:
```
# Linear Regression Using Statsmodels.api
import statsmodels.api as sm
import pandas as pd
```

2 Data Preprocessing

[3]:
```
# Reading the Data
data = pd.read_csv("insurance.csv", delimiter = ",")
```

[4]:
```
print(data.head(n=5))
```

```
   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
```

[11]:
```
# Getting the data ready for the models
data = pd.DataFrame(data)
predictors = data.iloc[:,:6] #all columns except charges
y = data.iloc[:,-1] #charges, response variable
df = data.copy() #copying the data

# Label encoding the data - from categorical to numerical
object_df = data.select_dtypes(include=['object']).copy()
object_df["sex"] = object_df["sex"].astype('category')
object_df["smoker"] = object_df["smoker"].astype('category')
object_df["region"] = object_df["region"].astype('category')

object_df["sex_binary"] = object_df["sex"].cat.codes
object_df["smoker_binary"] = object_df["smoker"].cat.codes
object_df["region_encoded"] = object_df["region"].cat.codes
```

```
#changing the columns in the data
df["sex"] = object_df["sex_binary"]
df["smoker"] = object_df["smoker_binary"]
df["region"] = object_df["region_encoded"]


print(df.head(n=5))
```

```
   age  sex     bmi  children  smoker  region      charges
0   19    0  27.900         0       1       3  16884.92400
1   18    1  33.770         1       0       2   1725.55230
2   28    1  33.000         3       0       2   4449.46200
3   33    1  22.705         0       0       1  21984.47061
4   32    1  28.880         0       0       1   3866.85520
```

3 Modeling

[13]:
```
# Models
predictors = df.iloc[:,:6] #all columns except charges - using the new␣
 ↪dataframe, 'df'
for i in predictors:
    X = predictors[i]
    X = sm.add_constant(X) #adding a constant - adding an intercept otherwise␣
 ↪we would get the wrong model
    model = sm.OLS(y,X)
    results = model.fit()
    print(f"{i} vs. charges: ", results.summary())
```

```
age vs. charges:                             OLS Regression Results
==============================================================================
Dep. Variable:                 charges   R-squared:                       0.089
Model:                             OLS   Adj. R-squared:                  0.089
Method:                  Least Squares   F-statistic:                     131.2
Date:                 Tue, 17 Nov 2020   Prob (F-statistic):           4.89e-29
Time:                         19:11:13   Log-Likelihood:                -14415.
No. Observations:                1338   AIC:                          2.883e+04
Df Residuals:                    1336   BIC:                          2.884e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3165.8850    937.149      3.378      0.001    1327.440    5004.330
age          257.7226     22.502     11.453      0.000     213.579     301.866
==============================================================================
Omnibus:                      399.600   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              864.239
```

```
Skew:                          1.733   Prob(JB):                    2.15e-188
Kurtosis:                      4.869   Cond. No.                          124.
==============================================================================

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

sex vs. charges: OLS Regression Results

Dep. Variable:	charges	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	4.400
Date:	Tue, 17 Nov 2020	Prob (F-statistic):	0.0361
Time:	19:11:13	Log-Likelihood:	-14475.
No. Observations:	1338	AIC:	2.895e+04
Df Residuals:	1336	BIC:	2.897e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.257e+04	470.072	26.740	0.000	1.16e+04	1.35e+04
sex	1387.1723	661.331	2.098	0.036	89.812	2684.532

Omnibus:	331.451	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	636.534
Skew:	1.496	Prob(JB):	6.00e-139
Kurtosis:	4.572	Cond. No.	2.63

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

bmi vs. charges: OLS Regression Results

Dep. Variable:	charges	R-squared:	0.039
Model:	OLS	Adj. R-squared:	0.039
Method:	Least Squares	F-statistic:	54.71
Date:	Tue, 17 Nov 2020	Prob (F-statistic):	2.46e-13
Time:	19:11:13	Log-Likelihood:	-14451.
No. Observations:	1338	AIC:	2.891e+04
Df Residuals:	1336	BIC:	2.892e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1192.9372	1664.802	0.717	0.474	-2072.974	4458.849

```
bmi          393.8730     53.251      7.397      0.000     289.409     498.337
==============================================================================
Omnibus:                      261.030   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              431.091
Skew:                           1.297   Prob(JB):                     2.45e-94
Kurtosis:                       4.004   Cond. No.                         160.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
children vs. charges:                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     6.206
Date:                Tue, 17 Nov 2020   Prob (F-statistic):             0.0129
Time:                        19:11:13   Log-Likelihood:                -14475.
No. Observations:                1338   AIC:                         2.895e+04
Df Residuals:                    1336   BIC:                         2.896e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.252e+04    446.450     28.049      0.000    1.16e+04    1.34e+04
children     683.0894    274.202      2.491      0.013     145.176    1221.002
==============================================================================
Omnibus:                      341.103   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              666.755
Skew:                           1.528   Prob(JB):                    1.64e-145
Kurtosis:                       4.619   Cond. No.                         2.65
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
smoker vs. charges:                              OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.619
Method:                 Least Squares   F-statistic:                     2178.
Date:                Tue, 17 Nov 2020   Prob (F-statistic):          8.27e-283
Time:                        19:11:13   Log-Likelihood:                -13831.
No. Observations:                1338   AIC:                         2.767e+04
Df Residuals:                    1336   BIC:                         2.768e+04
Df Model:                           1
Covariance Type:            nonrobust
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8434.2683    229.014     36.829      0.000    7985.002    8883.535
smoker      2.362e+04    506.075     46.665      0.000    2.26e+04    2.46e+04
==============================================================================
Omnibus:                      135.996   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              212.201
Skew:                           0.727   Prob(JB):                     8.34e-47
Kurtosis:                       4.300   Cond. No.                         2.60
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
region vs. charges:                            OLS Regression Results
==============================================================================
Dep. Variable:                 charges   R-squared:                       0.000
Model:                             OLS   Adj. R-squared:                 -0.001
Method:                  Least Squares   F-statistic:                    0.05149
Date:                 Tue, 17 Nov 2020   Prob (F-statistic):              0.821
Time:                         19:11:13   Log-Likelihood:                -14478.
No. Observations:                1338   AIC:                           2.896e+04
Df Residuals:                    1336   BIC:                           2.897e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.337e+04    562.359     23.781      0.000    1.23e+04    1.45e+04
region       -68.0449    299.858     -0.227      0.821    -656.289     520.199
==============================================================================
Omnibus:                      337.427   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              655.098
Skew:                           1.516   Prob(JB):                    5.59e-143
Kurtosis:                       4.600   Cond. No.                         3.83
==============================================================================

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