w2_assessment

August 29, 2021

1 Week 2 Python Assessment

This Jupyter Notebook is auxillary to the following assessment in this week. To complete this assessment, you will complete the 7 questions outlined in this document and use the output from your python cells as answers.

Your goal of this assignment is to construct regression and logistics models and interpret model paramters.

Run the following cell to initialize your environment and begin the assessment.

```
In [2]: #### RUN THIS

        import warnings
        warnings.filterwarnings('ignore')

        import numpy as np
        import statsmodels.api as sm
        import pandas as pd

        from sklearn.datasets import load_boston
        boston_dataset = load_boston()

        boston = pd.DataFrame(data=boston_dataset.data, columns=boston_dataset.feature_names)
        boston["MEDV"] = boston_dataset.target

        url = "nhanes_2015_2016.csv"
        NHANES = pd.read_csv(url)
        vars = ["BPXSY1", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "BMXBMI", "SMQ020"]
        NHANES = NHANES[vars].dropna()
        NHANES["smq"] = NHANES.SMQ020.replace({2: 0, 7: np.nan, 9: np.nan})
        NHANES["RIAGENDRx"] = NHANES.RIAGENDR.replace({1: "Male", 2: "Female"})
        NHANES["DMDEDUC2x"] = NHANES.DMDEDUC2.replace({1: "lt9", 2: "x9_11", 3: "HS", 4: "Some

        np.random.seed(123)
```

Now that your notebook is ready, begin answering the questions below.

1.0.1 Questions 1-3

The first three questions will be utilizing the Boston housing dataset seen in week 1.

Here is the description for each column:

- **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:** Nitric oxide concentration (parts per 10 million)
- **RM:** Average number of rooms per dwelling
- **AGE:** Proportion of owner-occupied units built prior to 1940
- **DIS:** Weighted distances to five Boston employment centers
- **RAD:** Index of accessibility to radial highways
- **TAX:** Full-value property tax rate per $\$10,000$
- **PTRATIO:** Pupil-teacher ratio by town
- **B:** $1000(Bk0.63)^2$, where Bk is the proportion of [people of African American descent] by town
- **LSTAT:** Percentage of lower status of the population
- **MEDV:** Median value of owner-occupied homes in $1000s

Uncomment and run the following code to generate a simple linear regression and output the model summary:

```
In [3]: model = sm.OLS.from_formula("MEDV ~ RM + CRIM", data=boston)
        result = model.fit()
        result.summary()
```

```
Out[3]: <class 'statsmodels.iolib.summary.Summary'>
        """
                                    OLS Regression Results
        ==============================================================================
        Dep. Variable:                   MEDV   R-squared:                       0.541
        Model:                            OLS   Adj. R-squared:                  0.539
        Method:                 Least Squares   F-statistic:                     295.9
        Date:                Tue, 30 Mar 2021   Prob (F-statistic):           1.15e-85
        Time:                        16:59:37   Log-Likelihood:                -1643.5
        No. Observations:                 506   AIC:                             3293.
        Df Residuals:                     503   BIC:                             3306.
        Df Model:                           2
        Covariance Type:            nonrobust
        ==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
        ------------------------------------------------------------------------------
        Intercept    -29.3017      2.592    -11.303      0.000     -34.395     -24.208
        RM             8.3975      0.406     20.706      0.000       7.601       9.194
        CRIM          -0.2618      0.033     -7.899      0.000      -0.327      -0.197
        ==============================================================================
        Omnibus:                      170.471   Durbin-Watson:                   0.805
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1034.461
Skew:                           1.331   Prob(JB):                    2.34e-225
Kurtosis:                       9.479   Cond. No.                         92.2
==============================================================================

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

Utilizing the above output, answer the following three questions:

Question 1 (You'll answer this question within the quiz that follows this notebook) What is the value of the coefficient for predictor **RM**?

Question 2 (You'll answer this question within the quiz that follows this notebook) Are the predictors for this model statistically significant, yes or no? (Hint: What are their p-values?)

Run the following code for question 3:

```
In [4]: ## For Question 3
        model = sm.OLS.from_formula("MEDV ~ RM + CRIM + LSTAT", data=boston)
        result = model.fit()
        result.summary()
```

```
Out[4]: <class 'statsmodels.iolib.summary.Summary'>
        """
                                    OLS Regression Results
        ==============================================================================
        Dep. Variable:                   MEDV   R-squared:                       0.646
        Model:                            OLS   Adj. R-squared:                  0.644
        Method:                 Least Squares   F-statistic:                     304.9
        Date:                Tue, 30 Mar 2021   Prob (F-statistic):          1.19e-112
        Time:                        17:00:55   Log-Likelihood:                -1577.8
        No. Observations:                 506   AIC:                             3164.
        Df Residuals:                     502   BIC:                             3180.
        Df Model:                           3
        Covariance Type:            nonrobust
        ==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
        ------------------------------------------------------------------------------
        Intercept     -2.4978      3.165     -0.789      0.430      -8.717       3.721
        RM             5.2092      0.442     11.785      0.000       4.341       6.078
        CRIM          -0.1011      0.032     -3.162      0.002      -0.164      -0.038
        LSTAT         -0.5804      0.048    -12.201      0.000      -0.674      -0.487
        ==============================================================================
        Omnibus:                      171.189   Durbin-Watson:                   0.822
        Prob(Omnibus):                  0.000   Jarque-Bera (JB):              623.248
        Skew:                           1.531   Prob(JB):                    4.61e-136
        Kurtosis:                       7.492   Cond. No.                         216.
```

```
==============================================================================

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

Question 3 (You'll answer this question within the quiz that follows this notebook) What happened to our R-Squared value when we added the third predictor **LSTAT** to our initial model?

Question 4 (You'll answer this question within the quiz that follows this notebook) What type of model should we use when our target outcome, or dependent variable is continuous?

1.0.2 Questions 5-6

The next two questions will involve the NHANES dataset.

Uncomment and run the following code to generate a logistics regression and output the model summary:

```
In [5]: model = sm.GLM.from_formula("smq ~ RIAGENDRx + RIDAGEYR + DMDEDUC2x", family=sm.famili
        result = model.fit()
        result.summary()
```

```
Out[5]: <class 'statsmodels.iolib.summary.Summary'>
        """
                         Generalized Linear Model Regression Results
        ==============================================================================
        Dep. Variable:                    smq   No. Observations:                 5093
        Model:                            GLM   Df Residuals:                     5086
        Model Family:                Binomial   Df Model:                            6
        Link Function:                  logit   Scale:                          1.0000
        Method:                          IRLS   Log-Likelihood:                -3201.2
        Date:                Tue, 30 Mar 2021   Deviance:                       6402.4
        Time:                        17:01:23   Pearson chi2:                 5.10e+03
        No. Iterations:                     4   Covariance Type:             nonrobust
        ======================================================================================
                                       coef    std err          z      P>|z|      [0.025
        --------------------------------------------------------------------------------------
        Intercept                   -2.3060      0.114    -20.174      0.000      -2.530
        RIAGENDRx[T.Male]            0.9096      0.060     15.118      0.000       0.792
        DMDEDUC2x[T.HS]              0.9434      0.090     10.521      0.000       0.768
        DMDEDUC2x[T.SomeCollege]     0.8322      0.084      9.865      0.000       0.667
        DMDEDUC2x[T.lt9]             0.2662      0.109      2.438      0.015       0.052
        DMDEDUC2x[T.x9_11]           1.0986      0.107     10.296      0.000       0.889
        RIDAGEYR                     0.0183      0.002     10.582      0.000       0.015
        ======================================================================================
        """
```

Question 5 (You'll answer this question within the quiz that follows this notebook) Which of our predictors has the largest coefficient?

Question 6 (You'll answer this question within the quiz that follows this notebook) Which values for DMDEDUC2x and RIAGENDRx are represented in our intercept, or what is our reference level?

Question 7 (You'll answer this question within the quiz that follows this notebook) What model should we use when our target outcome, or dependent variable is binary, or only has two outputs, 0 and 1.