Statsmodels and Plotly

Kristýna Němcová

Statsmodels

- statsmodels.org
- Python module
- statistical models, statistical tests and statistical data exploration
- using R-style formulas and pandas DataFrames

Patsy

- bringing R "formulas" to Python

```
y, X = dmatrices('gain_clf ~ clf + removed', data=df, return_type='dataframe')
X
```

	Intercept	clf[T.J48]	clf[T.JRip]	clf[T.LMT]	clf[T.Logistic]	clf[T.NaiveBayes]	clf[T.PART]	clf[T.RandomForest]	clf[T.SMO]	clf[T.SimpleLogistic]	removed
0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.0
2	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	5.0
4	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0
...	...	...	...	...	...	...	...	...	...	...	...
33502	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
33503	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
33504	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
33505	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
33506	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

33507 rows × 11 columns

Regression summary

- *Y = b1X1 + b2X2 + ... + A*
- intercept - all values 0
- t statistic - precision with which the coefficient was measured
- p-value - chance variable has no effect on the dependent variable

```
In [13]: mod = sm.OLS(y, X)      # Describe model

In [14]: res = mod.fit()         # Fit model

In [15]: print(res.summary())    # Summarize model
```

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.287
Method:                 Least Squares   F-statistic:                     6.636
Date:                Tue, 02 Feb 2021   Prob (F-statistic):           1.07e-05
Time:                        07:07:06   Log-Likelihood:                -375.30
No. Observations:                  85   AIC:                             764.6
Df Residuals:                      78   BIC:                             781.7
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      38.6517      9.456      4.087      0.000      19.826      57.478
Region[T.E]   -15.4278      9.727     -1.586      0.117     -34.793       3.938
Region[T.N]   -10.0170      9.260     -1.082      0.283     -28.453       8.419
Region[T.S]    -4.5483      7.279     -0.625      0.534     -19.039       9.943
Region[T.W]   -10.0913      7.196     -1.402      0.165     -24.418       4.235
Literacy       -0.1858      0.210     -0.886      0.378      -0.603       0.232
Wealth          0.4515      0.103      4.390      0.000       0.247       0.656
==============================================================================
Omnibus:                        3.049   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.694
Skew:                          -0.340   Prob(JB):                        0.260
Kurtosis:                       2.454   Cond. No.                         371.
```

Other

- `statsmodels.stats.multicomp.pairwise_tukeyhsd`
  - Tukey's post-hoc test

Plotly

- Open Source Graphing Libraries for Python, R, and JavaScript.
- licensed under the MIT license

- Scatter plots https://plotly.com/python/line-and-scatter/

Dash

- Plotly develops Dash
- Web interface and create dashboards with Plotly

- Interactive visualization https://dash.plotly.com/interactive-graphing
- Word2Vec https://dash-gallery.plotly.host/word2vec-enterprise/