MultivariateRegression

December 20, 2016

1 Multivariate Regression

Let's grab a small little data set of Blue Book car values:

```
In [1]: import pandas as pd

        df = pd.read_excel('http://cdn.sundog-soft.com/Udemy/DataScience/cars.xls')

In [2]: df.head()
```

Out[2]:

	Price	Mileage	Make	Model	Trim	Type	Cylinder	Liter	Doors	Cruise	Sound	Leather
0	17314.103129	8221	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	1
1	17542.036083	9135	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
2	16218.847862	13196	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
3	16336.913140	16342	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	0
4	16339.170324	19832	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	1

We can use pandas to split up this matrix into the feature vectors we're interested in, and the value we're trying to predict.

Note how we use pandas.Categorical to convert textual category data (model name) into an ordinal number that we can work with.

This is actually a questionable thing to do in the real world - doing a regression on categorical data only works well if there is some inherent order to the categories!

```
In [3]: import statsmodels.api as sm

        df['Model_ord'] = pd.Categorical(df.Model).codes
        X = df[['Mileage', 'Model_ord', 'Doors']]
        y = df[['Price']]
```