IDS 690: Unifying Data Science Interpreting Indicator Variables

Derek Wales and Prajwal Vijendra

In [1]:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
```

In [2]:

```
# Load in data
auto_df = pd.read_stata('automobile_dataset.dta')
auto_df.head()
```

Out[2]:

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio
0	AMC Concord	4099	22	3.0	2.5	11	2930	186	40	121	3.58
1	AMC Pacer	4749	17	3.0	3.0	11	3350	173	40	258	2.53
2	AMC Spirit	3799	22	NaN	3.0	12	2640	168	35	121	3.08
3	Buick Century	4816	20	3.0	4.5	16	3250	196	40	196	2.93
4	Buick Electra	7827	15	4.0	4.0	20	4080	222	43	350	2.41

Exercise 1

In [3]:

```
# Create new indicator variable
# 1 indicates mpg < 18
auto_df['guzzler'] = np.where(auto_df['mpg']<18,1,0)
```

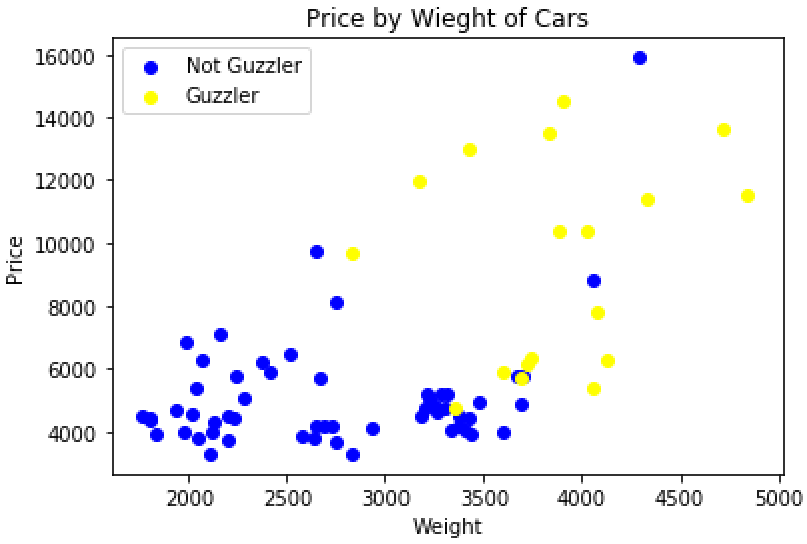
```
In [4]: # Linear regression on car price vs. guzzler
        smf.ols('price ~ C(guzzler)', data = auto_df).fit().summary()
```

```
Out[4]:
OLS Regression Results

      Dep. Variable:            price          R-squared:     0.379
             Model:              OLS      Adj. R-squared:     0.370
            Method:    Least Squares         F-statistic:     43.90
              Date: Fri, 21 Feb 2020  Prob (F-statistic):  5.38e-09
              Time:         13:13:11      Log-Likelihood:   -678.10
  No. Observations:               74                 AIC:     1360.
      Df Residuals:               72                 BIC:     1365.
          Df Model:                1
   Covariance Type:        nonrobust

                        coef   std err       t   P>|t|    [0.025     0.975]
        Intercept  5143.0893   312.807  16.442   0.000  4519.521   5766.658
  C(guzzler)[T.1]  4202.2440   634.243   6.626   0.000  2937.904   5466.584

        Omnibus:  37.244     Durbin-Watson:     1.348
  Prob(Omnibus):   0.000  Jarque-Bera (JB):   111.225
           Skew:   1.565          Prob(JB):  7.04e-25
       Kurtosis:   8.126          Cond. No.      2.50

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

Our simple linear regression shows that a gas guzzler adds roughly 4202.24 dollars to the price of a car.

Exercise 2

```
In [7]: # Scatterplot of price of cars by weight colored by guzzler
        plt.scatter(y='price', x='weight', label='Not Guzzler', color = 'blue', data=auto_df[auto_df.guzzler==0])
        plt.scatter(y='price', x='weight', label='Guzzler', color = 'yellow', data=auto_df[auto_df.guzzler==1])
        plt.ylabel('Price')
        plt.xlabel('Weight')
        plt.title('Price by Wieght of Cars')
        plt.legend()
        plt.show()
```



Given our analysis, we should control for weight in our regression. Not controling for weight would overestimate the coefficient for guzzler as most guzzlers tend to be more expensive than non-guzzlers.

Exercise 3

In [6]:
```
# Linear regression on car price vs. guzzler
# Additional controls
smf.ols('price ~ C(guzzler) + weight + C(foreign) + headroom + displacement',
data = auto_df).fit().summary()
```

Out[6]:
```
OLS Regression Results

      Dep. Variable:            price          R-squared:      0.596
             Model:               OLS     Adj. R-squared:      0.566
            Method:     Least Squares        F-statistic:      20.04
              Date:  Wed, 19 Feb 2020  Prob (F-statistic):   3.14e-12
              Time:          20:18:43     Log-Likelihood:    -662.20
  No. Observations:                74                AIC:      1336.
      Df Residuals:                68                BIC:      1350.
          Df Model:                 5
   Covariance Type:         nonrobust

                            coef    std err       t    P>|t|      [0.025     0.975]
            Intercept  -782.5353   1612.628  -0.485   0.629   -4000.484   2435.414
       C(guzzler)[T.1] 1977.1796    711.055   2.781   0.007     558.291   3396.068
C(foreign)[T.Foreign]  3278.9827    671.826   4.881   0.000    1938.375   4619.591
               weight     1.9634      0.702   2.797   0.007       0.563      3.364
             headroom  -736.7997    309.009  -2.384   0.020   -1353.418   -120.182
         displacement     8.9667      5.819   1.541   0.128      -2.646     20.579

        Omnibus:  22.179     Durbin-Watson:     1.409
  Prob(Omnibus):   0.000  Jarque-Bera (JB):    37.284
           Skew:   1.118          Prob(JB):  8.01e-09
       Kurtosis:   5.663          Cond. No.  2.36e+04


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.36e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Including additional control variable, our estimate of the guzzler coefficient is reduce to 1,977 from our previous estimate of 4,202. In other words, the additional cost incurred by buying a guzzler is reduced when controling for weight, foreign, headroom, and displacement.

Additionally, controling for guzzler-status, a foreign car costs 3,278 additional dollars on average, each additional lb of a car is associated with a 1.96 dollar increase in price, each unit of headroom is associated with a 736 dollar reduction in cost, and each unit of displacement is associated with a 8.96 dollar increase in cost.

Exercise 4

In [7]:

```
# Create indicator variables for repair record
auto_df = auto_df[pd.notnull(auto_df.rep78)].copy()
auto_df['rep_1'] = np.where(auto_df.rep78 == 1,1,0)
auto_df['rep_2'] = np.where(auto_df.rep78 == 2,1,0)
auto_df['rep_3'] = np.where(auto_df.rep78 == 3,1,0)
auto_df['rep_4'] = np.where(auto_df.rep78 == 4,1,0)
auto_df['rep_5'] = np.where(auto_df.rep78 == 5,1,0)
```

In [8]:

```
# Linear regression on car price vs. guzzler
# Control for repair record
smf.ols('price ~ rep_2 + rep_3 + rep_4 + rep_5 + weight + C(foreign) + headroom + displacement', data = auto_df).fit().summary() # C(guzzler)
```

Out[8]:

```
OLS Regression Results

   Dep. Variable:             price         R-squared:     0.562
           Model:               OLS    Adj. R-squared:     0.503
          Method:     Least Squares       F-statistic:     9.611
            Date:  Wed, 19 Feb 2020  Prob (F-statistic):  1.87e-08
            Time:          20:18:45    Log-Likelihood:   -619.34
No. Observations:                69               AIC:     1257.
    Df Residuals:                60               BIC:     1277.
        Df Model:                 8
 Covariance Type:         nonrobust

                             coef   std err       t  P>|t|    [0.025    0.975]
            Intercept  -3674.8333  2181.321  -1.685  0.097 -8038.125   688.458
C(foreign)[T.Foreign]   3565.2581   815.700   4.371  0.000  1933.616  5196.901
                rep_2   1292.4864  1717.908   0.752  0.455 -2143.841  4728.814
                rep_3   1546.1189  1582.091   0.977  0.332 -1618.534  4710.771
                rep_4   1319.9236  1649.062   0.800  0.427 -1978.692  4618.539
                rep_5   1917.3066  1732.508   1.107  0.273 -1548.226  5382.839
               weight      2.1325     0.890   2.396  0.020     0.352     3.913
             headroom   -750.7992   351.685  -2.135  0.037 -1454.274   -47.325
         displacement     15.4064     7.517   2.049  0.045     0.369    30.443

      Omnibus:  16.791     Durbin-Watson:     1.414
Prob(Omnibus):   0.000  Jarque-Bera (JB):    19.847
         Skew:   1.131          Prob(JB):  4.90e-05
     Kurtosis:   4.335          Cond. No.  4.44e+04


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.44e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

According to our analysis, buying a car with an acceptable repair record costs 1,546 dollars more, on average, compared to simialar car (similar weight, headroom, displacement, and place of manufacturing) with a very poor repair record.

Additionally, controling for repair record, a foreign car costs 3,565 additional dollars on average, each additional lb of a car is associated with a 2.13 dollar increase in price, each unit of headroom is associated with a 750 dollar reduction in cost, and each unit of displacement is associated with a 15 dollar increase in cost.

Exercise 5

In [9]:

```
# Linear regression on car price vs. guzzler
# Interaction effects for foreign and guzzler
smf.ols('price ~ weight + C(foreign) + guzzler + C(foreign)*guzzler + headroom
+ displacement', data = auto_df).fit().summary() # C(guzzler)
```

Out[9]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.598
Model:	OLS	**Adj. R-squared:**	0.560
Method:	Least Squares	**F-statistic:**	15.40
Date:	Wed, 19 Feb 2020	**Prob (F-statistic):**	1.02e-10
Time:	20:18:46	**Log-Likelihood:**	-616.32
No. Observations:	69	**AIC:**	1247.
Df Residuals:	62	**BIC:**	1262.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-491.1286	1806.853	-0.272	0.787	-4102.977	3120.720
C(foreign)[T.Foreign]	3064.1019	717.185	4.272	0.000	1630.469	4497.735
weight	1.4155	0.872	1.623	0.110	-0.328	3.159
guzzler	1234.8723	786.967	1.569	0.122	-338.253	2807.997
C(foreign)[T.Foreign]:guzzler	2367.5358	1635.982	1.447	0.153	-902.745	5637.817
headroom	-672.3043	314.885	-2.135	0.037	-1301.751	-42.857
displacement	15.4387	7.492	2.061	0.044	0.463	30.415

Omnibus:	24.422	**Durbin-Watson:**	1.484
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	40.778
Skew:	1.300	**Prob(JB):**	1.40e-09
Kurtosis:	5.725	**Cond. No.**	2.58e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.58e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

According to our analysis, there are additional costs associated with buying a guzzler and buying foreign cars (1,234 and 3,064 respectively). However, there if your car is both a guzzler and foreign, there is an additional 2,367 cost for a car that is both a foreign guzzler on top all previous costs.

Exercise 6

In [10]:

```
foreign = 3064.1019
guzzler = 1234.8723
foreign_x_guzzler = 2367.5358
```

In [11]:

```
# Difference between foreign guzzler and foreign non-guzzler
(foreign + guzzler + foreign_x_guzzler) - (foreign)
```

Out[11]:

```
3602.4081
```

Exercise 7

In [12]:

```
# Difference between domestic non-guzzler and foreign non-guzzler
0 - (foreign)
```

Out[12]:

```
-3064.1019
```

Exercise 8

In [13]:

```
# Linear regression on car price vs. guzzler
# Interaction effects for foreign and guzzler
smf.ols('price ~ C(foreign)*mpg + headroom + weight + displacement', data = auto_df).fit().summary() # C(guzzler)
```

Out[13]:

```
OLS Regression Results

     Dep. Variable:             price           R-squared:     0.586
            Model:                OLS      Adj. R-squared:     0.546
           Method:      Least Squares         F-statistic:     14.63
             Date:   Wed, 19 Feb 2020  Prob (F-statistic):  2.54e-10
             Time:           20:18:48      Log-Likelihood:   -617.37
 No. Observations:                 69                 AIC:     1249.
     Df Residuals:                 62                 BIC:     1264.
         Df Model:                  6
  Covariance Type:          nonrobust

                                  coef   std err       t   P>|t|     [0.025      0.975]
                  Intercept  -1.113e+04  4726.813  -2.355   0.022  -2.06e+04   -1684.462
       C(foreign)[T.Foreign]  1.072e+04  3113.992   3.444   0.001   4498.775    1.69e+04
                        mpg   236.0976   114.808   2.056   0.044      6.599     465.596
   C(foreign)[T.Foreign]:mpg  -273.1404   120.012  -2.276   0.026   -513.042     -33.239
                   headroom  -456.5527   332.445  -1.373   0.175  -1121.101     207.996
                     weight     2.9201     1.002   2.914   0.005      0.917       4.923
               displacement    18.2165     7.228   2.520   0.014      3.769      32.664

       Omnibus:  21.421     Durbin-Watson:     1.499
 Prob(Omnibus):   0.000  Jarque-Bera (JB):    30.381
          Skew:   1.249          Prob(JB):  2.53e-07
      Kurtosis:   5.079          Cond. No.  7.12e+04


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.12e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

According to our analysis, each additional mpg a domestic car has, the price increases on average 236 dollars per mile. If a car is foreign, the price increases by 1,072 on average and the with each additional mpg the average car price decreases by 37 dollars.