Interpreting Indicator Variables

Xiaoquan Liu & Yuanjing Zhu

In [ ]:
```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
%config InlineBackend.figure_format = 'retina'
```

In [ ]:
```
df = pd.read_stata('https://github.com/nickeubank/MIDS_Data/blob'
                   '/master/automobile_dataset.dta?raw=true')
df.head()
```

Out[ ]:

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

Exercise 1

create a new variable named guzzler.

In [ ]:
```
# create a new column called guzzler
df['guzzler'] = np.where(df['mpg'] < 18, 1, 0)

# regress price on guzzler
mod_1 = smf.ols('price ~ guzzler', data=df).fit()
mod_1.summary()
```

Out[ ]:

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:	Thu, 23 Feb 2023	**Prob (F-statistic):**	5.38e-09
Time:	20:14:05	**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
guzzler	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

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

Holding all other variables constant, a guzzler car will be $4202.24 more expensive than a non-guzzler car.

Exercise 2

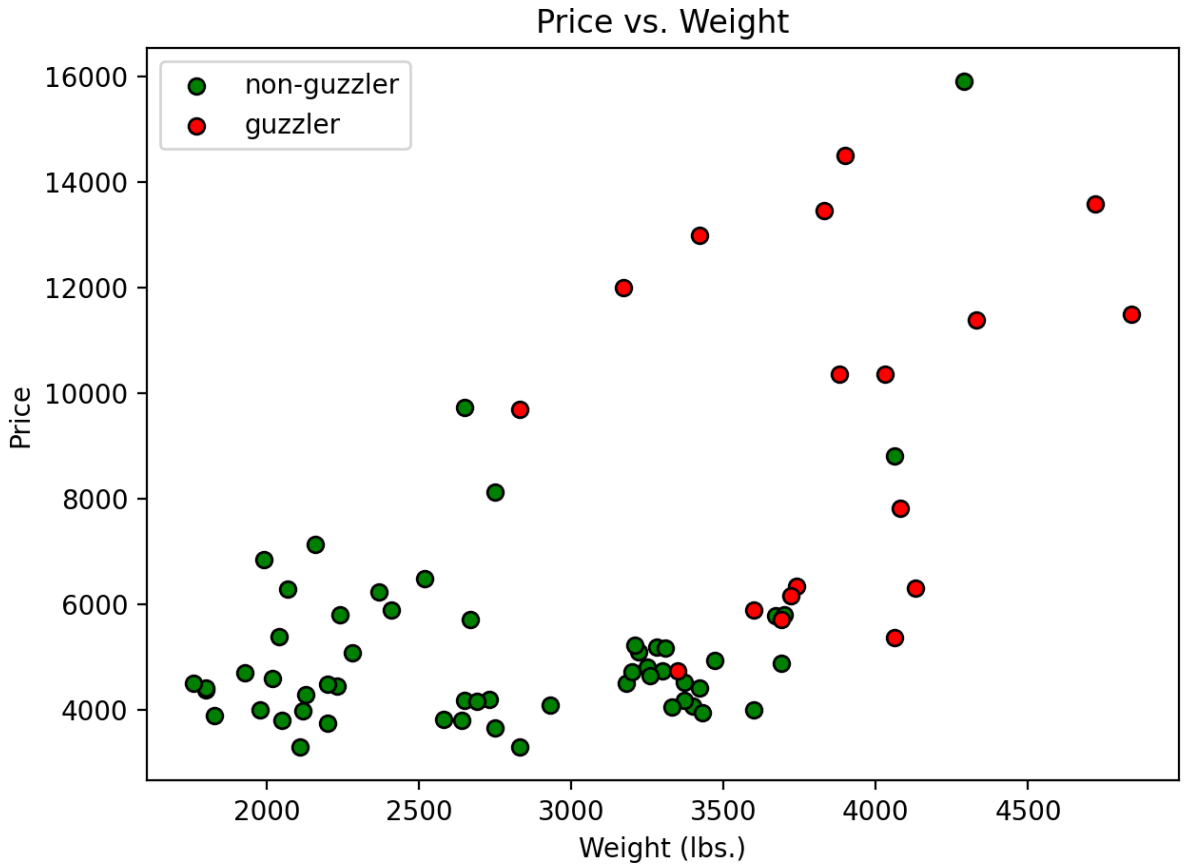
Create a scatter plot of price against weight

In [ ]:

```
price_guzzler = df.loc[df['guzzler'] == 1, 'price']
weight_guzzler = df.loc[df['guzzler'] == 1, 'weight']
price_non_guzzler = df.loc[df['guzzler'] == 0, 'price']
weight_non_guzzler = df.loc[df['guzzler'] == 0, 'weight']
weight = df.loc[:, 'weight']

# plot the data
plt.figure(figsize=(7, 5), dpi=100)
plt.scatter(weight_non_guzzler, price_non_guzzler, color='green', \
            label='non-guzzler', edgecolors='black')
plt.scatter(weight_guzzler, price_guzzler, color='red', \
            label='guzzler', edgecolors='black')
plt.xlabel('Weight (lbs.)')
plt.ylabel('Price')
plt.title('Price vs. Weight')
```

```
plt.legend()
plt.show()
```



From the scatter plot, we can see that guzzler cars are generally heavier and more expensive than non-guzzler cars.

Not controlling for weight might lead to omitted variable bias in the regression in Exercise 1 and the **direction of the bias is positive** (higher than real).

This is because if we don't include weight in the regression, the positive effect of weight on price will be included into the effect of guzzler on price. In other words, we would be overestimating the effect of guzzler on price. Therefore, the direction of the bias would be positive, meaning that the coefficient estimate for guzzler would be higher than real.

Exercise 3

Regress price on guzzler, weight, foreign, headroom, and displacement

In [ ]:
```
# regress price on guzzler, weight, foreign, headroom, and displacement.
mod_3 = smf.ols('price ~ guzzler + weight + foreign + headroom + displacement', \
                data=df).fit()
mod_3.summary()
```

Out[ ]:

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:	Thu, 23 Feb 2023	**Prob (F-statistic):**	3.14e-12
Time:	20:14:06	**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
foreign[T.Foreign]	3278.9827	671.826	4.881	0.000	1938.375	4619.591
guzzler	1977.1796	711.055	2.781	0.007	558.291	3396.068
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

Notes:
[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.

Coefficient interpretation:

- Holding all other variables constant, a foreign car will be $3278.98 more expensive than a domestic car.
- Holding all other variables constant, a guzzler car will be $1977.18 more expensive than a non-guzzler car.
- Holding all other variables constant, 1 lb increase in weight will lead to $1.96 increase in price.
- Holding all other variables constant, 1 inch increase in headroom will lead to $736.80 decrease in price.
- Holding all other variables constant, 1 cubic inch increase in displacement will lead to $8.97 increase in price.

After we control on weight, the coefficient estimate for guzzler decreases from 4202 to 1977. This confirms our prediction in Q3 that the coefficient estimate for guzzler in Q1 is overestimated. The inclusion of weight and other variables helps to explain some of the variation in price that was previously being attributed to guzzler.

Exercise 4

Create five separate indicator variables from rep78 and regress price on indicators for values 2 through 5.

In [ ]:
```
# regress price on 4 indicators, headroom, weight, foreign, and displacement
mod_4 = smf.ols("price ~ C(rep78) + headroom "
                "+ weight + foreign + displacement", data=df).fit()
mod_4.summary()
```

Out[ ]:

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:	Thu, 23 Feb 2023	**Prob (F-statistic):**	1.87e-08
Time:	20:14:06	**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(rep78)[T.2.0]	1292.4864	1717.908	0.752	0.455	-2143.841	4728.814
C(rep78)[T.3.0]	1546.1189	1582.091	0.977	0.332	-1618.534	4710.771
C(rep78)[T.4.0]	1319.9236	1649.062	0.800	0.427	-1978.692	4618.539
C(rep78)[T.5.0]	1917.3066	1732.508	1.107	0.273	-1548.226	5382.839
foreign[T.Foreign]	3565.2581	815.700	4.371	0.000	1933.616	5196.901
headroom	-750.7992	351.685	-2.135	0.037	-1454.274	-47.325
weight	2.1325	0.890	2.396	0.020	0.352	3.913
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

Notes:
[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.

The coefficient of C(rep78)[T.3.0] is 1546, which means that a car with acceptable repair record will be $1546 more expensive than a car with very poor repair record, holding all other variables constant.

In [ ]:
```
df['rep78'].value_counts()
```

Out[ ]:
```
3.0    30
4.0    18
5.0    11
2.0     8
1.0     2
Name: rep78, dtype: int64
```

In [ ]:
```
mode_try = smf.ols("price ~ C(rep78) + headroom + weight + foreign + displacement", \
                   data=df).fit()
mode_try.summary()
```

Out[ ]:

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:	Thu, 23 Feb 2023	**Prob (F-statistic):**	1.87e-08
Time:	20:14:06	**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(rep78)[T.2.0]	1292.4864	1717.908	0.752	0.455	-2143.841	4728.814
C(rep78)[T.3.0]	1546.1189	1582.091	0.977	0.332	-1618.534	4710.771
C(rep78)[T.4.0]	1319.9236	1649.062	0.800	0.427	-1978.692	4618.539
C(rep78)[T.5.0]	1917.3066	1732.508	1.107	0.273	-1548.226	5382.839
foreign[T.Foreign]	3565.2581	815.700	4.371	0.000	1933.616	5196.901
headroom	-750.7992	351.685	-2.135	0.037	-1454.274	-47.325
weight	2.1325	0.890	2.396	0.020	0.352	3.913
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

Notes:
[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.

Exercise 5

Regress price on guzzler, foreign and their interaction, controlling for headroom, weight and displacement.

In [ ]:
```
# regress price on guzzler, foreign and their interaction
mod_5 = smf.ols("price ~ guzzler*foreign + headroom + weight + displacement", \
                data=df).fit()
mod_5.summary()
```

Out[ ]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.619
Model:	OLS	**Adj. R-squared:**	0.585
Method:	Least Squares	**F-statistic:**	18.15
Date:	Thu, 23 Feb 2023	**Prob (F-statistic):**	2.21e-12
Time:	20:14:06	**Log-Likelihood:**	-660.00
No. Observations:	74	**AIC:**	1334.
Df Residuals:	67	**BIC:**	1350.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-391.7038	1588.834	-0.247	0.806	-3563.030	2779.622
foreign[T.Foreign]	2929.3402	679.319	4.312	0.000	1573.413	4285.267
guzzler	1354.9011	760.244	1.782	0.079	-162.552	2872.354
guzzler:foreign[T.Foreign]	2797.6787	1381.501	2.025	0.047	40.190	5555.167
headroom	-736.8717	302.195	-2.438	0.017	-1340.056	-133.688
weight	1.6417	0.705	2.330	0.023	0.235	3.048
displacement	12.6296	5.972	2.115	0.038	0.710	24.549

Omnibus:	26.353	**Durbin-Watson:**	1.421
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	46.874
Skew:	1.311	**Prob(JB):**	6.63e-11
Kurtosis:	5.885	**Cond. No.**	2.39e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.39e+04. This might indicate that there are strong multicollinearity or other numerical problems.

The coefficient of the interaction term is 2797.68, which means that the difference of how guzzler affects foreign and domestic cars is 2797.68. In other words, the price difference between foreign and

domestic guzzler cars is $(2929 + 2797.68)$, while the price difference between foreign and domestic non-guzzler cars is 2929. Therefore the extent to which the effect of guzzler on price differs depending on whether the car is foreign or domestic is the coefficient of the interaction term, which is 2797.68.

Exercise 6

In [ ]:
```
price_diff = mod_5.params['guzzler'] + mod_5.params['guzzler:foreign[T.Foreign]']
print(f"The price difference between a foreign guzzler and a foreign non-guzzler is \
      {price_diff:.2f}")
```

```
The price difference between a foreign guzzler and a foreign non-guzzler is 4152.58
```

Exercise 7

In [ ]:
```
price_diff = mod_5.params['foreign[T.Foreign]']
print(f"The price difference between a domestic non-guzzler and a foreign non-guzzler is \
      {price_diff:.2f}")
```

```
The price difference between a domestic non-guzzler and a foreign non-guzzler is 2929.34
```

Exercise 8

Regress price on foreign, mpg and their interaction, controlling for headroom, weight and displacement.

In [ ]:
```
# regress price on foreign, mpg and their interaction
mod_8 = smf.ols("price ~ foreign*mpg + headroom + weight + displacement", \
                data=df).fit()
mod_8.summary()
```

Out[ ]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.599
Model:	OLS	**Adj. R-squared:**	0.564
Method:	Least Squares	**F-statistic:**	16.71
Date:	Thu, 23 Feb 2023	**Prob (F-statistic):**	1.12e-11
Time:	20:14:06	**Log-Likelihood:**	-661.86
No. Observations:	74	**AIC:**	1338.
Df Residuals:	67	**BIC:**	1354.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.232e+04	4465.992	-2.758	0.007	-2.12e+04	-3404.206
foreign[T.Foreign]	1.176e+04	2796.011	4.208	0.000	6184.000	1.73e+04
mpg	259.8139	109.998	2.362	0.021	40.257	479.371
foreign[T.Foreign]:mpg	-314.4806	109.360	-2.876	0.005	-532.764	-96.197
headroom	-484.5821	319.958	-1.515	0.135	-1123.222	154.058
weight	3.4327	0.856	4.008	0.000	1.723	5.142
displacement	14.4670	5.839	2.478	0.016	2.813	26.121

Omnibus:	22.563	**Durbin-Watson:**	1.442
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	33.595
Skew:	1.228	**Prob(JB):**	5.07e-08
Kurtosis:	5.204	**Cond. No.**	6.89e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.89e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Coefficient interpretation of the main independent variables:

- Holding all other variables constant, a foreign car will be $11760 more expensive than a domestic car when their mileage equals zero.
- Holding all other variables constant, 1 mpg increase in mileage will lead to $259.81 increase in price for domestic car.
- Holding all other variables constant, 1 inch increase in headroom will lead to $484.58 decrease in price.
- Holding all other variables constant, 1 lb increase in weight will lead to $3.43 increase in price.
- Holding all other variables constant, 1 cubic inch increase in displacement will lead to $14.47 increase in price.

Coefficient interpretation on the interaction term:

- Holding all other variables constant, as domestic car gets 1 mpg increase in mileage, the price will increase $259.81. (mpg: 259.8139)
- Holding all other variables constant, as foreign car gets 1 mpg increase in mileage, the price change will be $314.48 less than that of a domestic car. (foreign[T.Foreign]:mpg: -314.4806)
- In other words, holding all other variables constant, as foreign car gets 1 mpg increase in mileage, the price will decrease $54.67.$(259.81-314.48 = 54.67$)