In [1]:
```
# Note that the answers to each of these questions should be the direct result of running
# appropriate Python or R code and not involve any manual processing of dataset files. Answers
# without either the code or results will not receive any grade.
# 1. For the next exercise, you are going to use the "airline_costs.csv" dataset.
# The dataset has the following attributes:
# i. Airline name
# ii. Length of flight in miles
# iii. Speed of plane in miles per hour
# iv. Daily flight time per plane in hours
# v. Customers served in 1000s
# vi. Total operating cost in cents per revenue ton-mile
# vii. Revenue in tons per aircraft mile
# viii. Ton-mile load factor
# ix. Available capacity
# x. Total assets in $100,000s
# xi. Investments and special funds in $100,000s
# xii. Adjusted assets in $100,000s
# (Implement this exercise in Python language; import 'pandas', 'statsmodels.api' libraries)
# Use a linear regression model to predict the number of customers each airline serves
# from its length of flight and daily flight time per plane. Next, build another regression
# model to predict the total assets of an airline from the customers served by the airline.
# Do you have any insight about the data from the last two regression models? (20 points)
```

In [5]:
```
#import data
import pandas as pd
import statsmodels.api as sm
df = pd.read_csv('airline_costs.csv')
df.head()
```

Out[5]:

	Airline	FlightLength	PlaneSpeed	DailyFlightTime	CustomersServed	TotalOperatingCost	Revenue	LoadFactor	AvailableCapacity	T
0	All-American	57	133	6.10	20200	116.3	0.96	0.400	2.400	
1	American	270	216	6.93	56928	43.0	3.98	0.689	5.776	
2	Bonanza	100	140	4.45	183	141.5	0.79	0.358	2.207	
3	Braniff	176	182	6.60	11869	50.6	2.57	0.557	4.614	
4	Capital	142	167	7.47	41097	51.0	2.68	0.510	5.255	

In [8]:
```
#model 1: length of flight and daily fight time
y = df["CustomersServed"]
x = df[["FlightLength","DailyFlightTime"]]
lr_model = sm.OLS(endog=y,exog=x).fit()

#model 2: total assets ~ customers
y = df["TotalAssets"]
x = df["CustomersServed"]
lr_model_2 = sm.OLS(endog=y,exog=x).fit()
```

In [7]:
```
lr_model.summary()
```

Out[7]:

OLS Regression Results

Dep. Variable:	CustomersServed	**R-squared (uncentered):**	0.779
Model:	OLS	**Adj. R-squared (uncentered):**	0.764
Method:	Least Squares	**F-statistic:**	51.18
Date:	Sun, 06 Dec 2020	**Prob (F-statistic):**	3.07e-10
Time:	21:25:01	**Log-Likelihood:**	-330.53
No. Observations:	31	**AIC:**	665.1
Df Residuals:	29	**BIC:**	667.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
FlightLength	185.0209	30.284	6.109	0.000	123.083	246.959
DailyFlightTime	-1390.1041	674.170	-2.062	0.048	-2768.937	-11.271

Omnibus:	5.022	**Durbin-Watson:**	1.958
Prob(Omnibus):	0.081	**Jarque-Bera (JB):**	3.780
Skew:	0.843	**Prob(JB):**	0.151
Kurtosis:	3.294	**Cond. No.**	52.0

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [10]:
```
lr_model_2.summary()
```

Out[10]:

OLS Regression Results

Dep. Variable:	TotalAssets	**R-squared (uncentered):**	0.833
Model:	OLS	**Adj. R-squared (uncentered):**	0.827
Method:	Least Squares	**F-statistic:**	149.6
Date:	Sun, 06 Dec 2020	**Prob (F-statistic):**	3.45e-13
Time:	21:25:14	**Log-Likelihood:**	-205.69
No. Observations:	31	**AIC:**	413.4
Df Residuals:	30	**BIC:**	414.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CustomersServed	0.0187	0.002	12.233	0.000	0.016	0.022

Omnibus:	4.660	**Durbin-Watson:**	1.605
Prob(Omnibus):	0.097	**Jarque-Bera (JB):**	4.053
Skew:	-0.253	**Prob(JB):**	0.132
Kurtosis:	4.697	**Cond. No.**	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Our takeaway from the first model is that the number of customers is positively associated with larger flight lengths and negatively associated with higher flight times. This seems counterintuitive as more customers would be mean more flights, and therefore more flight time. However, this is accounted for by the fact that the explanatory power of this variable is very diminished. Overly long flights do, however, require a tradeoff in quality of service, and therefore may decrease customers.

Larger flight lengths imply that the airline handles larger distances and therefore may be an international airline or a cross-domestic airline as opposed to a local or municipal service provided, and thus handles a larger customer base by necessity.

The second model shows that the number of customers served is positively associated with the total assets of the airline with a strong correlation (>0.80). Given that both models have high enough R squared values, we canmake the argument that the total assets of a company are then also positively associated with larger flight lengths, and to a lesser extent, negatively associated with higher flight times.