Tutorial #1 (part 2)

Fundamentals of Analytics (Simple and Multiple Regression)

Objective:

This is the second part of the tutorial #1.

The objective of this tutorial is to familiarize participants with simple and multiple linear regression and interpretation of the results.

This tutorial will cover the following topics:

1. Simple linear regression
2. Multiple linear regression

Tools: Jupyter notebooks, Python with the following libraries: pandas, matplotlib, statsmodels.

Prerequisites: Basic Python knowledge and familiarity with descriptive statistics.

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats as smt
```

In [2]:

```
# to read from a locally saved file use this
# df = pd.read_csv("Davis-weight.csv")

# to read directly from github, use the following URL path
df = pd.read_csv("https://raw.githubusercontent.com/agrianalytics/fundamentals/master/w
eight.csv")
```

How many rows and columns does the data set have?

In [3]:

```
df.shape
```

Out[3]:

```
(200, 6)
```

In [4]:

```
df.head()
```

Out[4]:

	ID	sex	weight	height	repwt	repht
0	1	M	77	182	77.0	180.0
1	2	F	58	161	51.0	159.0
2	3	F	53	161	54.0	158.0
3	4	M	68	177	70.0	175.0
4	5	F	59	157	59.0	155.0

In [5]:

```
df.dtypes
```

Out[5]:

```
ID          int64
sex        object
weight      int64
height      int64
repwt     float64
repht     float64
dtype: object
```

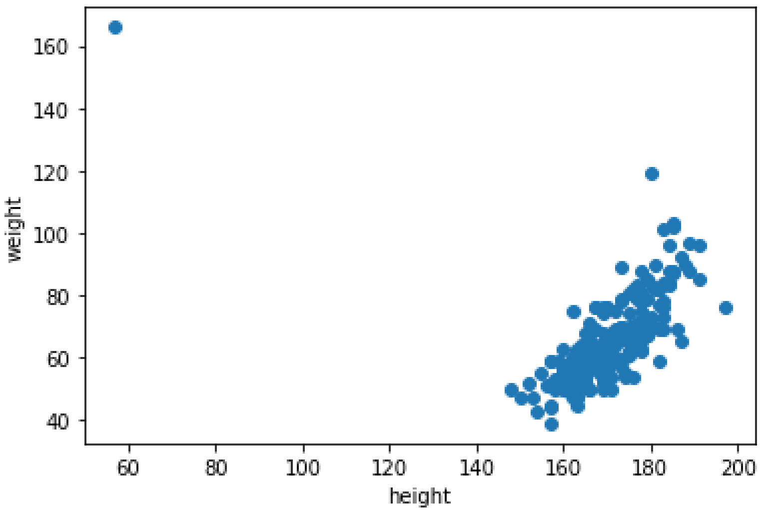
Create a scatter plot of height and weight. What do you notice?

In [6]:

```
plt.scatter(x=df.height, y=df.weight)
plt.xlabel('height')
plt.ylabel('weight')
```

Out[6]:

```
Text(0, 0.5, 'weight')
```



We remove the outlier.

In [7]:

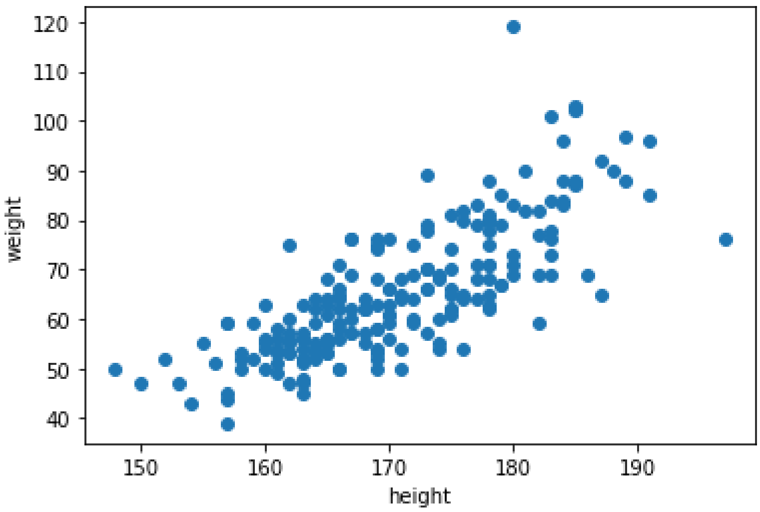
```
df = df[df.height > 80]
```

In [8]:

```
plt.scatter(x=df.height, y=df.weight)
plt.xlabel('height')
plt.ylabel('weight')
```

Out[8]:

```
Text(0, 0.5, 'weight')
```



We will use the **statsmodel.formula.api** to specify a model. We will then call the *fit()* method of the model and then print out the results.

In [9]:

```
model = smf.ols(formula='weight ~ height', data = df)
result = model.fit()
print(result.summary())
```

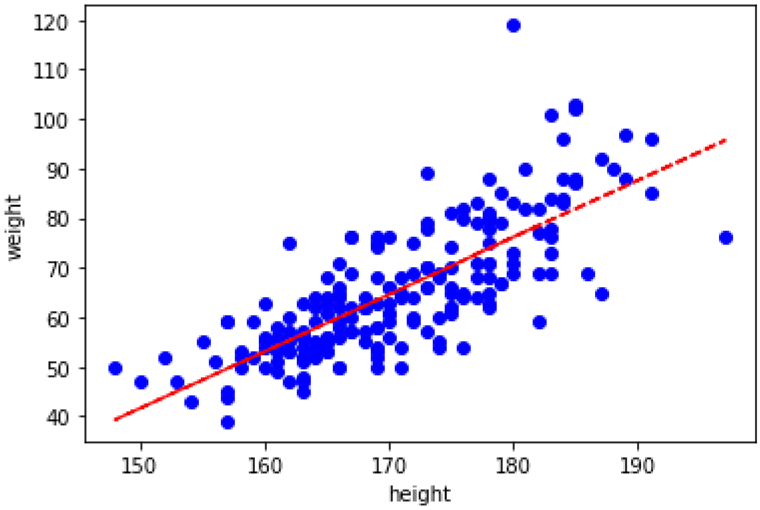
```
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                 weight   R-squared:
0.594
Model:                            OLS   Adj. R-squared:
0.592
Method:                 Least Squares   F-statistic:                     2
88.3
Date:                Sun, 09 Jun 2019   Prob (F-statistic):           2.01
e-40
Time:                        16:21:22   Log-Likelihood:                -70
7.79
No. Observations:                 199   AIC:                             1
420.
Df Residuals:                     197   BIC:                             1
426.
Df Model:                           1
Covariance Type:            nonrobust
==========================================================================
====
                coef    std err          t      P>|t|      [0.025      0.
975]
--------------------------------------------------------------------------
----
Intercept   -130.7470     11.563    -11.308      0.000    -153.550    -10
7.944
height         1.1492      0.068     16.978      0.000       1.016
1.283
==========================================================================
====
Omnibus:                       33.873   Durbin-Watson:
1.844
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7
7.622
Skew:                           0.766   Prob(JB):                     1.40
e-17
Kurtosis:                       5.648   Cond. No.                     3.27
e+03
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The condition number is large, 3.27e+03. This might indicate that ther
e are
strong multicollinearity or other numerical problems.
```

We can plot the actual and fitted values on a graph.

'bo' and 'r--' are formatting strings for the markers. ('bo' = blue circles and 'r--' = red dashes)

In [10]:

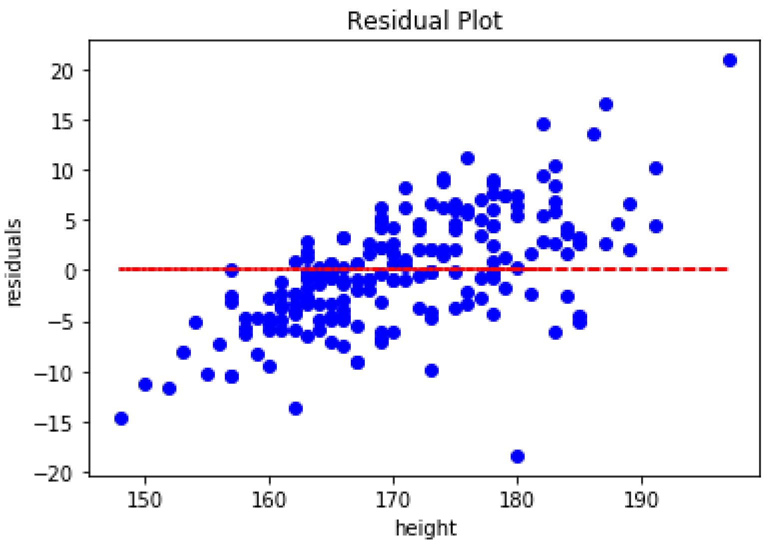
```
plt.plot(df.height, df.weight, 'bo')
plt.plot(df.height, result.fittedvalues, 'r--')
plt.xlabel('height')
plt.ylabel('weight')
plt.show()
```



Residual plots are drawn as part of the diagnostics of the regression model. What does this one tell us?

In [17]:

```
plt.plot(df.height, result.resid, 'bo')
plt.plot(df.height, [0]*len(df.height), 'r--')
plt.xlabel('height')
plt.ylabel('residuals')
plt.title('Residual Plot')
plt.show()
```



Answer the following questions:

1. How much of the variability in weight can be explained by this model?
2. What is the relationship between height and weight? Write down an equation.
3. Is *height* statistically significant in explaining weight?
4. What can we do to make the model better?

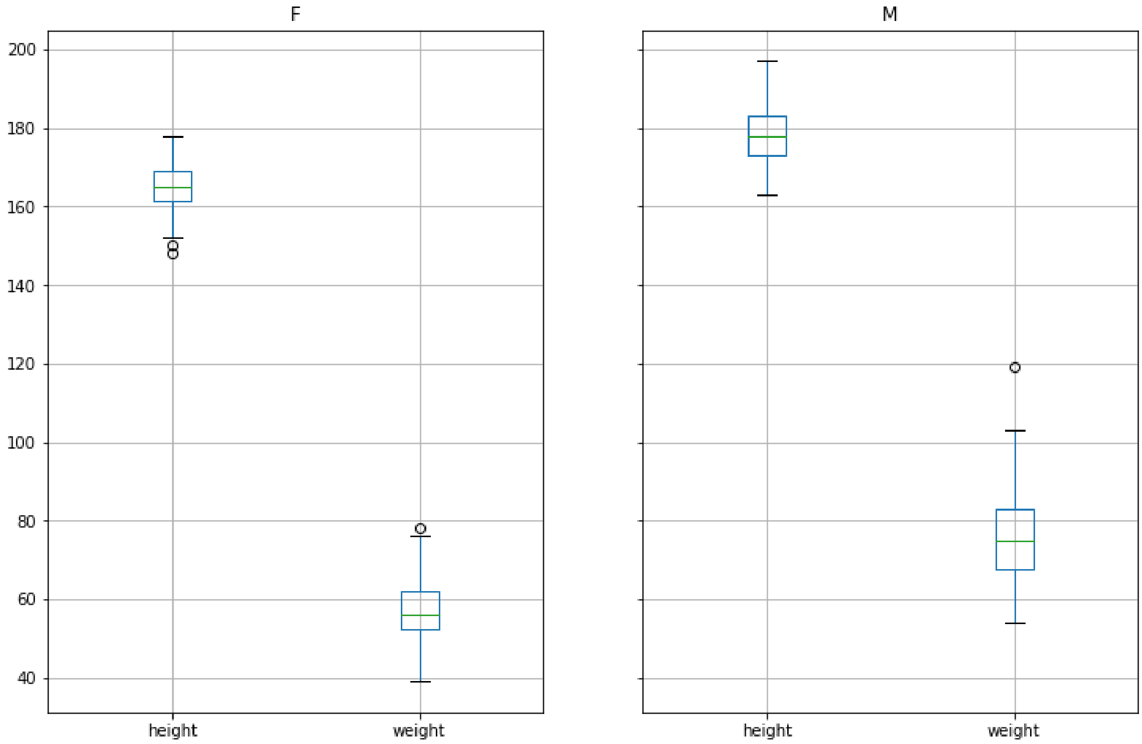
Could other variables be important?

In [12]:

```
df[['sex', 'height', 'weight']].groupby('sex').boxplot(figsize=(12, 8))
```

Out[12]:

```
F         AxesSubplot(0.1,0.15;0.363636x0.75)
M    AxesSubplot(0.536364,0.15;0.363636x0.75)
dtype: object
```



Does gender matter? Let's add another variable 'sex' to the model.

In [13]:

```
model = smf.ols(formula='weight ~ height + sex', data = df)
result = model.fit()
print(result.summary())
```

```
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                 weight   R-squared:
0.636
Model:                            OLS   Adj. R-squared:
0.633
Method:                 Least Squares   F-statistic:                     1
71.6
Date:                Sun, 09 Jun 2019   Prob (F-statistic):           8.54
e-44
Time:                        16:21:23   Log-Likelihood:                -69
6.80
No. Observations:                 199   AIC:                             1
400.
Df Residuals:                     196   BIC:                             1
409.
Df Model:                           2
Covariance Type:            nonrobust
==========================================================================
====
                 coef    std err          t      P>|t|      [0.025      0.
975]
--------------------------------------------------------------------------
----
Intercept    -76.6362     15.755     -4.864      0.000    -107.708      -4
5.564
sex[T.M]       8.2162      1.717      4.784      0.000       4.830       1
1.603
height         0.8107      0.096      8.485      0.000       0.622
0.999
==========================================================================
====
Omnibus:                       38.506   Durbin-Watson:
1.859
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               10
0.338
Skew:                           0.822   Prob(JB):                     1.63
e-22
Kurtosis:                       6.066   Cond. No.                     4.71
e+03
==========================================================================
====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified.
[2] The condition number is large, 4.71e+03. This might indicate that ther
e are
strong multicollinearity or other numerical problems.
```

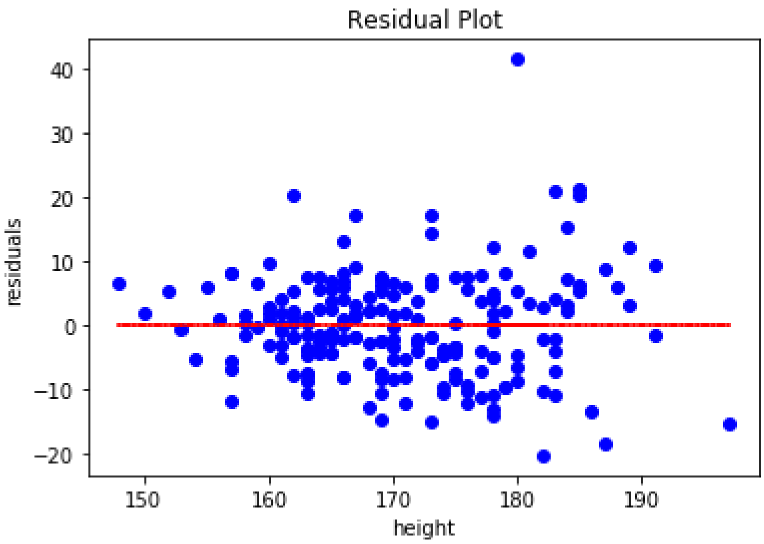
Note that sex[T.M] *(i.e. Treatment M)* is a **dummy** variable.

Answer the following questions:

1. Is this a beter model than the previous one?
2. How much of the variability in weight can be explained by this model?
3. What is the relationship between the dependent and independent variables? Write down an equation.
4. Which variables are statistically significant in explaining weight? What are the null and alternative hypotheses in this case?

In [14]:

```
plt.plot(df.height, result.resid, 'bo')
plt.plot(df.height, [0]*len(df.height), 'r--')
plt.xlabel('height')
plt.ylabel('residuals')
plt.title('Residual Plot')
plt.show()
```



In [15]:

```
aov = smt.anova.anova_lm(result)
aov
```

Out[15]:

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	17730.725290	17730.725290	271.177649	8.010402e-39
height	1.0	4707.485084	4707.485084	71.997322	5.238355e-15
Residual	196.0	12815.297164	65.384169	NaN	NaN

In the simple model, what if we interchange height and weight? Is there **causality**?

In [16]:

```
model = smf.ols(formula='height ~ weight', data = df)
result = model.fit()
print(result.summary())
```

```
                            OLS Regression Results
==========================================================================
====
Dep. Variable:                 height   R-squared:
0.594
Model:                            OLS   Adj. R-squared:
0.592
Method:                 Least Squares   F-statistic:                     2
88.3
Date:                Sun, 09 Jun 2019   Prob (F-statistic):           2.01
e-40
Time:                        16:21:23   Log-Likelihood:                -62
8.29
No. Observations:                 199   AIC:                             1
261.
Df Residuals:                     197   BIC:                             1
267.
Df Model:                           1
Covariance Type:            nonrobust
==========================================================================
====
                 coef    std err          t      P>|t|      [0.025      0.
975]
--------------------------------------------------------------------------
----
Intercept    136.8366      2.029     67.446      0.000     132.836     14
0.838
weight         0.5169      0.030     16.978      0.000       0.457
0.577
==========================================================================
====
Omnibus:                        5.915   Durbin-Watson:
1.945
Prob(Omnibus):                  0.052   Jarque-Bera (JB):
7.807
Skew:                           0.170   Prob(JB):                       0.
0202
Kurtosis:                       3.909   Cond. No.
334.
==========================================================================
====

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