ASSIGNMENT 5

1. Run linear regression fit and plot Y vs X (scatter plot) and a linear fitted line

2. Answer the following:

- Is a linear model appropriate? Please explain.
- Are outliers present? IF there are outliers, do you expect them to be influential? Why?

3. Perform an analysis of linear regression:

- Do a graphical analysis: a. plot histogram of residuals; b. plot residuals vs. predictor
- Are the requiremtns for linear regression met?

a. Linearity: The data should show a linear trend. If there is a nonlinear trend an advanced regression method from another book or later course should be applied.

b. Nearly normal residuals: Generally the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points.

c. Constant variability. The variability of points around the least squares line remains roughly constant.

d. Independent observations. Be cautious about applying regression to time series data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis.

In [1]:
```
import pandas as pd
import statsmodels.formula.api as sm
import statsmodels.graphics
import matplotlib.pyplot as plt
pd.options.display.max_rows = 6
```

In [2]:
```
mod6=pd.ExcelFile('Module6_Exercise.xlsx')
```

Set 1

In [30]:
```
#Read Set 1 worksheet from Module 6 csv.
df1=pd.read_excel(mod6, 'Set 1')
df1
```

Out[30]:

	y	x
0	38.858144	7.266278
1	40.891148	7.985333
2	48.971648	9.387120
...	...	...
97	39.739810	7.612336
98	7.963448	1.227335
99	46.095461	9.545883

100 rows × 2 columns

In [31]:
```
#Regression analysis of Set 1
result1 = sm.ols(formula='df1.y ~ df1.x', data=df1).fit()
result1.summary()
```

Out[31]:

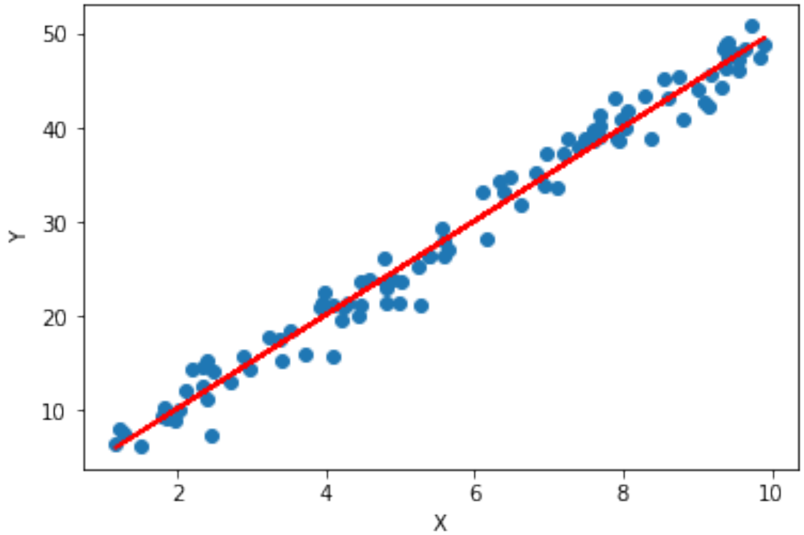
OLS Regression Results

Dep. Variable:	df1.y	**R-squared:**	0.979
Model:	OLS	**Adj. R-squared:**	0.979
Method:	Least Squares	**F-statistic:**	4579.
Date:	Tue, 03 Jul 2018	**Prob (F-statistic):**	4.47e-84
Time:	14:53:27	**Log-Likelihood:**	-206.03
No. Observations:	100	**AIC:**	416.1
Df Residuals:	98	**BIC:**	421.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.2381	0.469	0.508	0.613	-0.693	1.169
df1.x	4.9843	0.074	67.669	0.000	4.838	5.130

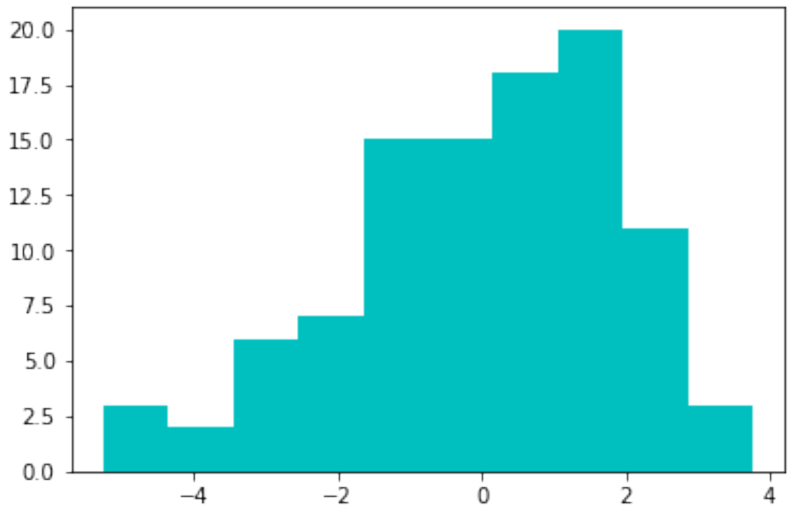
Omnibus:	4.971	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.083	**Jarque-Bera (JB):**	4.783
Skew:	-0.536	**Prob(JB):**	0.0915
Kurtosis:	2.988	**Cond. No.**	15.9

In [5]:
```
#Plot Y vs X (scatter plot) and a linear fitted line
plt.plot(df1.x, df1.y, 'o')
intercept, slope = result1.params
plt.plot(df1.x, intercept + slope * df1.x , 'r-', label='Fitted Line')
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```

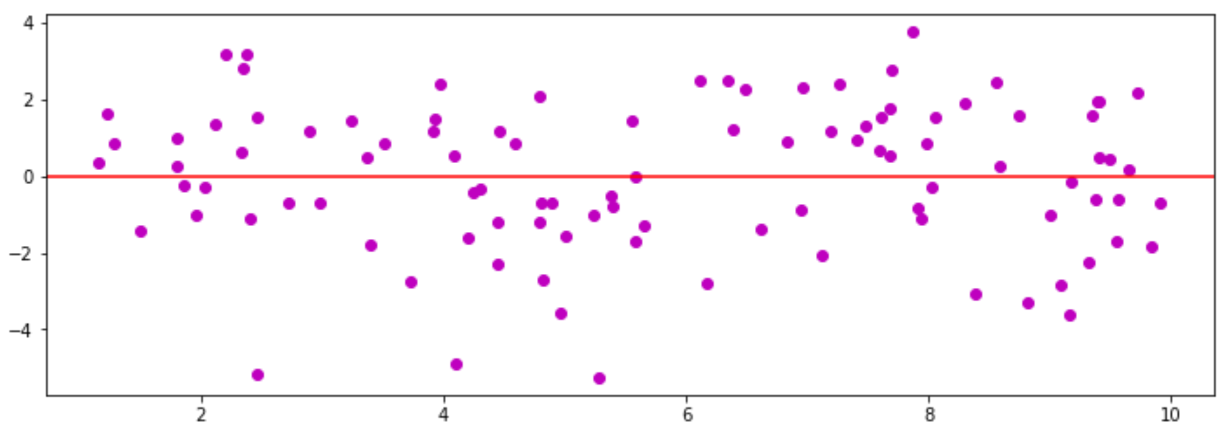


The plot above shows a relatively strong upward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. There aren't any noticeable outliers. The strengh of the linear fit is explained by the R-squared value 0.979.

In [6]:
```
#Histogram of residuals
residual1=(df1.y-result1.predict(df1.x))
residual1
plt.hist(residual1,bins=10,color='c')
plt.show()
```



In [7]:
```
#Plot Residuals vs predictor
plt.figure(figsize=(12,4))
plt.plot(df1.x, residual1,'o',color='m')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
```



a. Linearity: As mentioned above, the scatter & line plot above confirms linearity of the dataset with a R-squared value of 0.979.

b. Nearly Normal Residuals: The residuals are slightly left skewed but show a nearly normal distribution.

c. Constant variability: The variability of residuals around the least squares line remains roughly constant with larger values of x. Most residuals lie within the range -4 and 4.

d. Independent Observations: We assume observations are independent. (We don't have additional information of the data determine whether or not observations are independent.)

Set 2

In [32]:
```
#Read Set 2 worksheet from Module 6csv.
df2=pd.read_excel(mod6, 'Set 2')
result2 = sm.ols(formula='df2.y ~ df2.x', data=df2).fit()
result2.summary()
```

Out[32]:

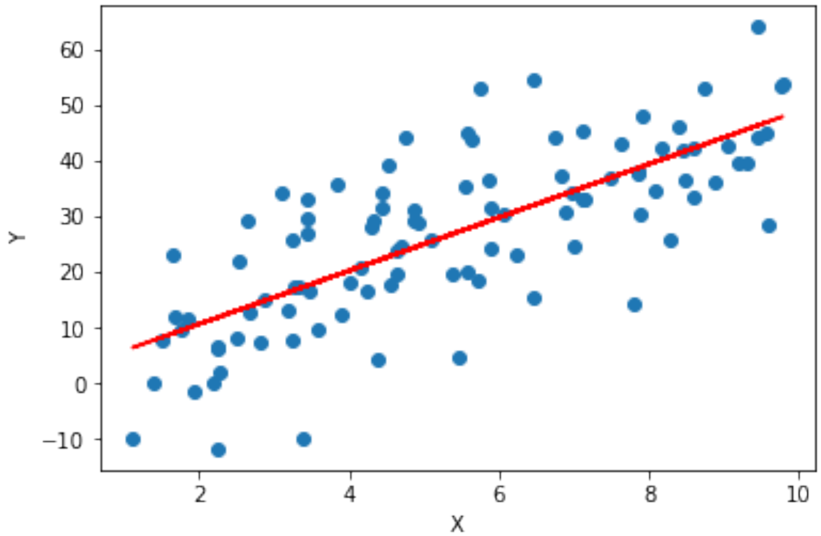
OLS Regression Results

Dep. Variable:	df2.y	**R-squared:**	0.555
Model:	OLS	**Adj. R-squared:**	0.551
Method:	Least Squares	**F-statistic:**	122.4
Date:	Tue, 03 Jul 2018	**Prob (F-statistic):**	6.11e-19
Time:	14:53:42	**Log-Likelihood:**	-375.73
No. Observations:	100	**AIC:**	755.5
Df Residuals:	98	**BIC:**	760.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0956	2.547	0.430	0.668	-3.958	6.149
df2.x	4.7774	0.432	11.062	0.000	3.920	5.634

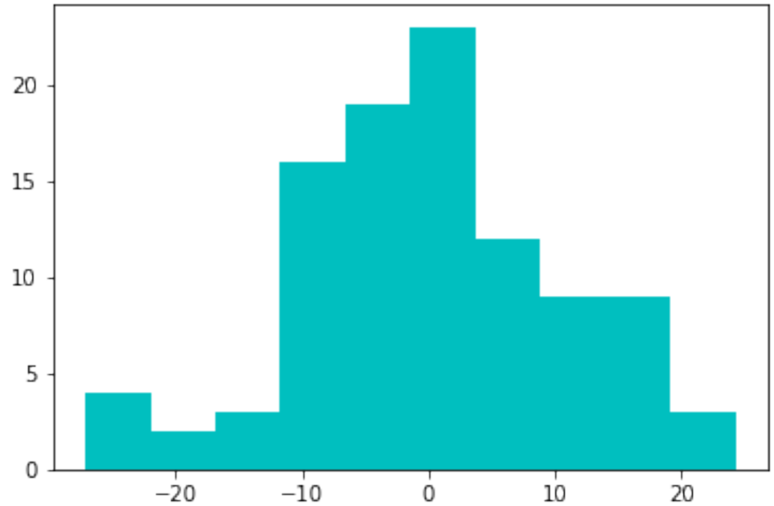
Omnibus:	0.254	**Durbin-Watson:**	2.043
Prob(Omnibus):	0.881	**Jarque-Bera (JB):**	0.079
Skew:	-0.065	**Prob(JB):**	0.961
Kurtosis:	3.045	**Cond. No.**	14.7

In [9]:
```
#Plot Y vs X (scatter plot) and a linear fitted line
plt.plot(df2.x, df2.y, 'o')
result2.params
intercept, slope = result2.params
plt.plot(df2.x, intercept + slope * df2.x , 'r-', label='Fitted Line')
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```

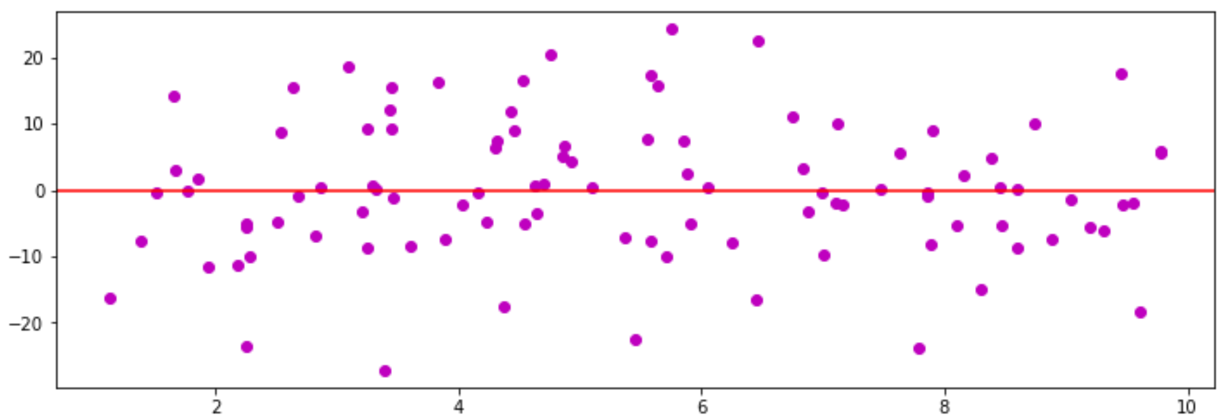


The plot above shows an slightly upward linear trend that, while evident, is not as strong as the previous plot (Set 1) and the R-squared value is 0.555 which supports this statement. The residual variability in the data around the line is more apparent relative to the strength of the relationship between x and y compared to Set1. There seems to be no apparent outliers that are influential points.

In [10]:
```
#Histogram or residuals
residual2=(df2.y-result2.predict(df2.x))
plt.hist(residual2,bins=10,color='c')
plt.show()
```



In [11]:
```
#Plot Residuals vs predictor
plt.figure(figsize=(12,4))
plt.plot(df2.x, residual2,'o',color='m')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
```



a. Linearity: As mentioned above, the dataset appears to have a linear relationship, but not as strong as Set 1, with a R-squared value of 0.555.

b. Nearly Normal Residuals: The residuals show a unimodal, nearly normal distribution.

c. Constant variability: The variability of residuals around the least squares line remains roughly constant throughout the predictor variable X. Most residuals lie within the range [-20, 20] and this also supports that the linear relationship is not as strong as Set 1.

d. Independent Observations: We assume observations are independent. (We don't have additional information of the data determine whether or not observations are independent.)

Set 3

In [12]:

```
#Read Set 3 worksheet from Module 6
df3=pd.read_excel(mod6, 'Set 3')
result3=sm.ols('df3.y~df3.x', data=df3).fit()
result3.summary()
```

Out[12]:

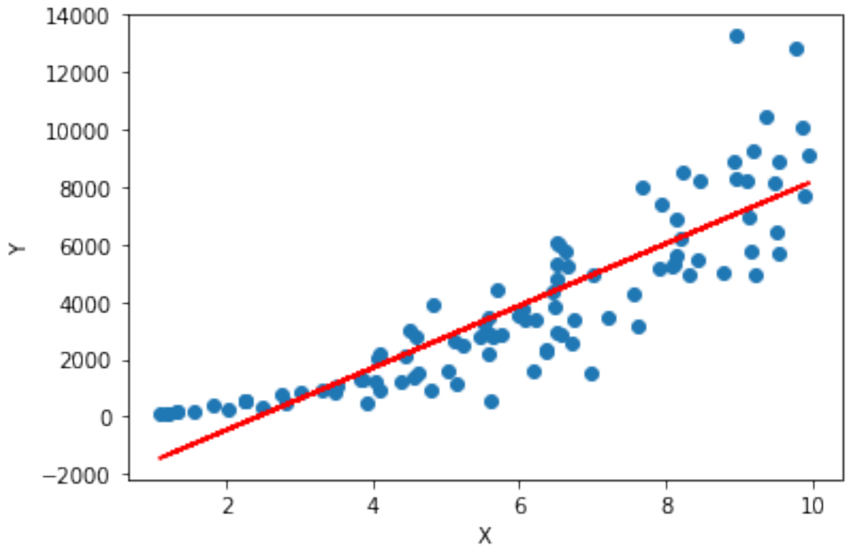
OLS Regression Results

Dep. Variable:	df3.y	**R-squared:**	0.755
Model:	OLS	**Adj. R-squared:**	0.753
Method:	Least Squares	**F-statistic:**	302.4
Date:	Tue, 03 Jul 2018	**Prob (F-statistic):**	1.04e-31
Time:	14:27:22	**Log-Likelihood:**	-873.07
No. Observations:	100	**AIC:**	1750.
Df Residuals:	98	**BIC:**	1755.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-2636.1748	402.741	-6.546	0.000	-3435.400	-1836.949
df3.x	1081.8266	62.216	17.388	0.000	958.361	1205.292

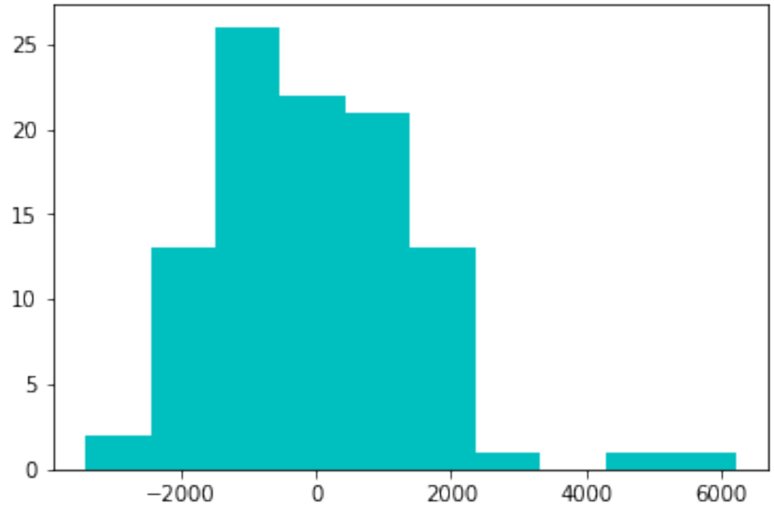
Omnibus:	21.170	**Durbin-Watson:**	2.159
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	37.896
Skew:	0.863	**Prob(JB):**	5.90e-09
Kurtosis:	5.474	**Cond. No.**	17.6

In [13]:
```
#Plot Y vs X (scatter plot) and a linear fitted line
plt.plot(df3.x, df3.y, 'o')
intercept, slope = result3.params
plt.plot(df3.x, intercept + slope * df3.x , 'r-', label='Fitted Line')
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```

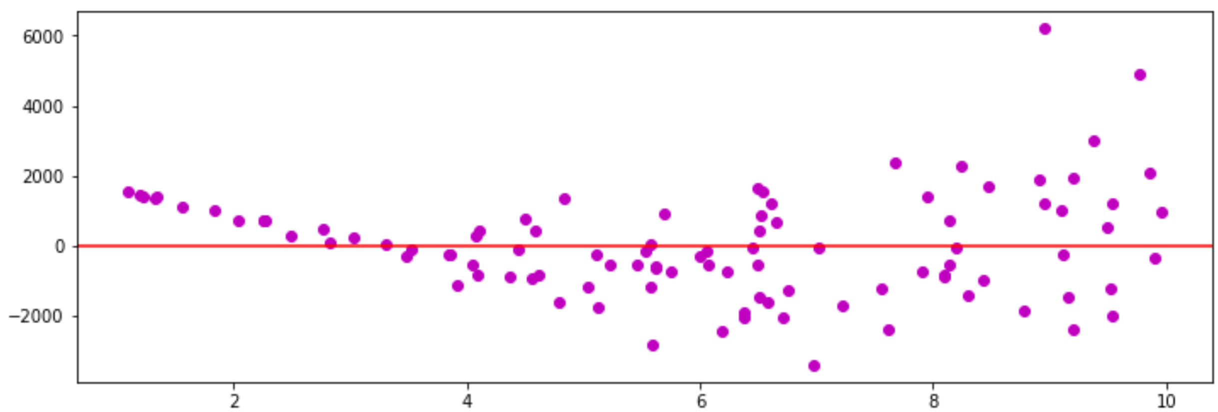


By looking at the graph above, we can see that a linear model is not quite appropriate. It seems to have more of an exponential relationship. This is a good example of when R-squared value (0.755) is closer to 1, indicating a strong postive relationship, but not by a linear relationship. There are two points between 8 < x < 10 that could be outliers but they do not greatly affect the slope of the line so they aren't too influential.

In [14]:
```
#Histogram of residuals
residual3=(df3.y-result3.predict(df3.x))
plt.hist(residual3,bins=10,color='c')
plt.show()
```



In [15]:
```
#Plot Residuals vs predictor
plt.figure(figsize=(12,4))
plt.plot(df3.x, residual3,'o',color='m')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
```



a. Linearity: As mentioned above, this dataset can't be modeled with a linear relationship.

b. Nearly Normal Residuals: The residuals is unimodal and shows a nearly normal distribution with a right tail.

c. Constant variability: The variability of the data around the line increases with larger values of x.

d. Independent Observations: We assume observations are independent. (We don't have additional information of the data determine whether or not observations are independent.)

Set 4

In [33]:
```
#Read Set 4 worksheet from Module 6
df4=pd.read_excel(mod6, 'Set 4')
result4=sm.ols('df4.y~df4.x', data=df4).fit()
result4.summary()
```

Out[33]:

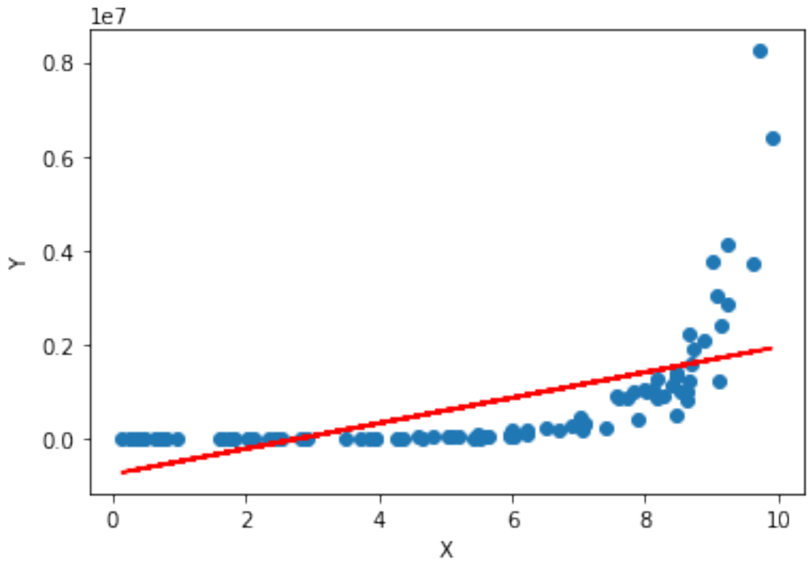
OLS Regression Results

Dep. Variable:	df4.y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.373
Method:	Least Squares	**F-statistic:**	59.97
Date:	Tue, 03 Jul 2018	**Prob (F-statistic):**	8.87e-12
Time:	14:54:50	**Log-Likelihood:**	-1526.2
No. Observations:	100	**AIC:**	3056.
Df Residuals:	98	**BIC:**	3062.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.535e+05	2.1e+05	-3.585	0.001	-1.17e+06	-3.36e+05
df4.x	2.707e+05	3.49e+04	7.744	0.000	2.01e+05	3.4e+05

Omnibus:	102.143	**Durbin-Watson:**	2.077
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1253.666
Skew:	3.381	**Prob(JB):**	5.89e-273
Kurtosis:	18.973	**Cond. No.**	12.4

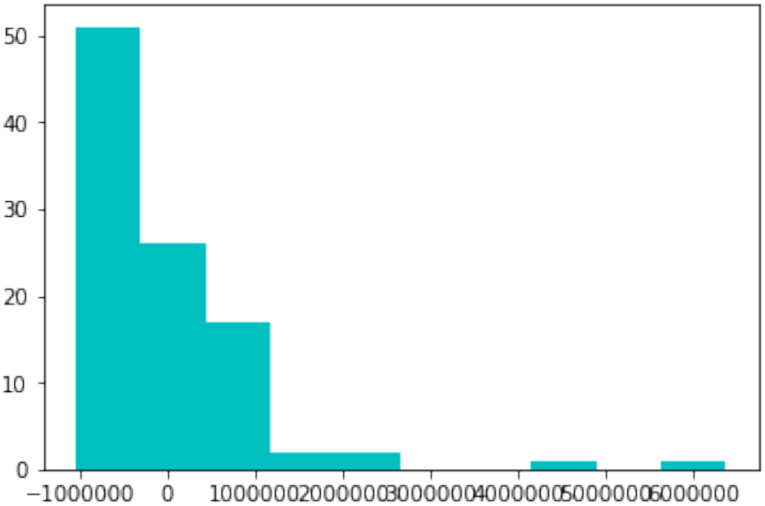
In [17]:
```
#Plot Y vs X (scatter plot) and a linear fitted line
plt.plot(df4.x, df4.y, 'o')
intercept, slope = result4.params
plt.plot(df4.x, intercept + slope * df4.x , 'r-', label='Fitted Line')
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```



By looking at the graph above, we can see that a linear model is not ppropriate. It definitely seems to have a stronger exponential-like relationship. There is a strong relationship between the variables. However, the correlation is not very strong (R-squared=0.380), and the relationship is not linear. The two points on the right seem like outliers and they aren't close to the linear line, which suggests they could be influencing the slope. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line. However, in this case even if we didn't have those two points, since this isnt a linear relationship, we can't draw a linear line so it doesn't really affect the slope.

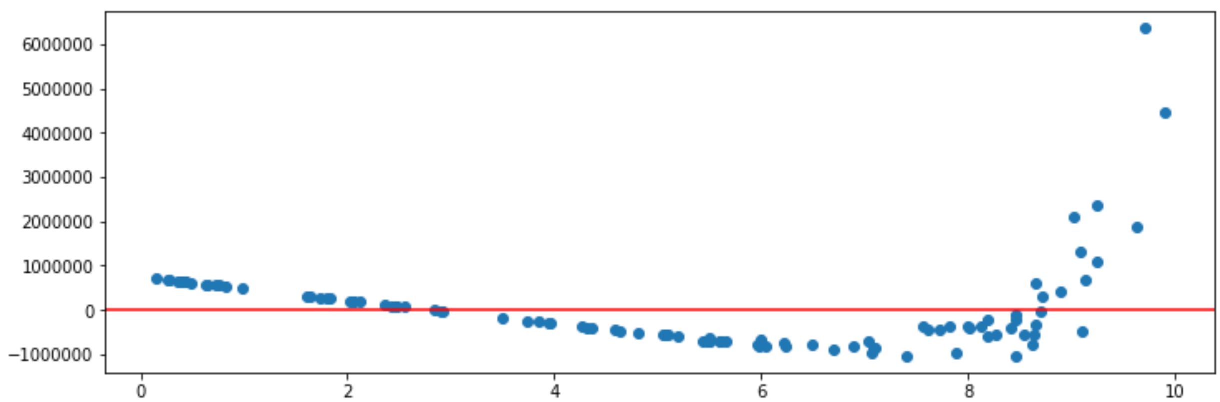
In [18]:

```
#Histogram of residuals
residual4=(df4.y-result4.predict(df4.x))
plt.hist(residual4,bins=10, color='c')
plt.show()
```



In [19]:

```
#Plot Residuals vs predictor
plt.figure(figsize=(12,4))
plt.plot(df4.x, residual4,'o')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
```



a. Linearity: As mentioned above, it is not a linear relationship.

b. Nearly Normal Residuals:The residuals do not show a normal distribution.

c. Constant variability: The variability of residuals isn't constant: Begins with positive residuals till when X=3 then shows only negative residuals till X=9 and back to positive residuals.

d. Independent Observations: We assume observations are independent. (We don't have additional information of the data determine whether or not observations are independent.)

Set 5

In [20]:
```
#Read Set 5 worksheet from Module 6
df5=pd.read_excel(mod6, 'Set 5')
```

In [21]:
```
result5=sm.ols('df5.y~df5.x', data=df5).fit()
result5.summary()
```

Out[21]:

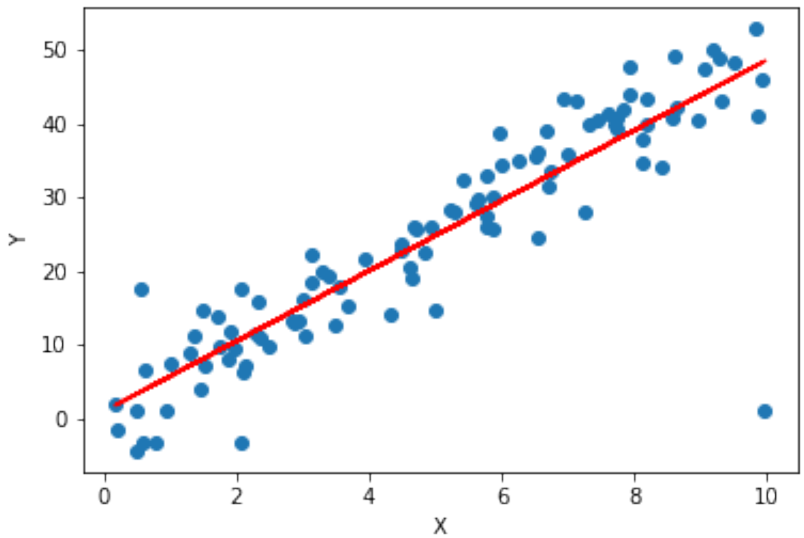
OLS Regression Results

Dep. Variable:	df5.y	**R-squared:**	0.806
Model:	OLS	**Adj. R-squared:**	0.804
Method:	Least Squares	**F-statistic:**	411.9
Date:	Tue, 03 Jul 2018	**Prob (F-statistic):**	4.70e-37
Time:	14:27:23	**Log-Likelihood:**	-334.42
No. Observations:	101	**AIC:**	672.8
Df Residuals:	99	**BIC:**	678.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.9213	1.346	0.685	0.495	-1.749	3.591
df5.x	4.7671	0.235	20.294	0.000	4.301	5.233

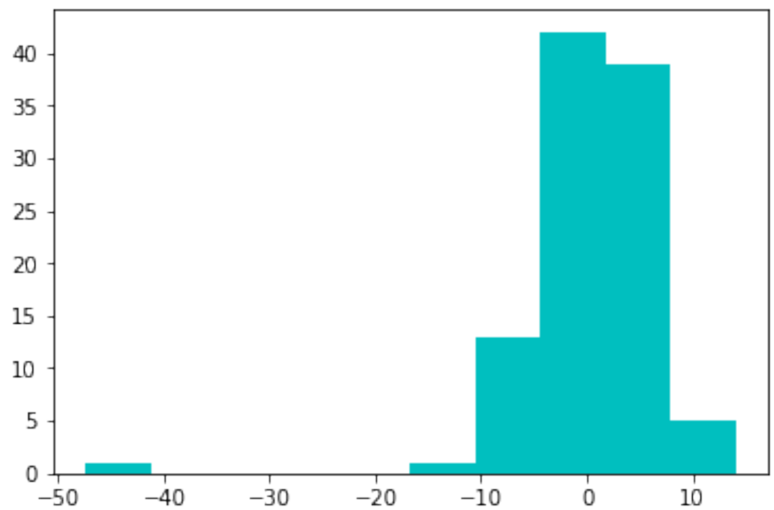
Omnibus:	113.783	**Durbin-Watson:**	1.491
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2578.951
Skew:	-3.591	**Prob(JB):**	0.00
Kurtosis:	26.691	**Cond. No.**	11.8

In [22]:
```
#Plot Y vs X (scatter plot) and a linear fitted line
plt.plot(df5.x, df5.y, 'o')
intercept, slope = result5.params
plt.plot(df5.x, intercept + slope * df5.x , 'r-', label='Fitted Line')
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```

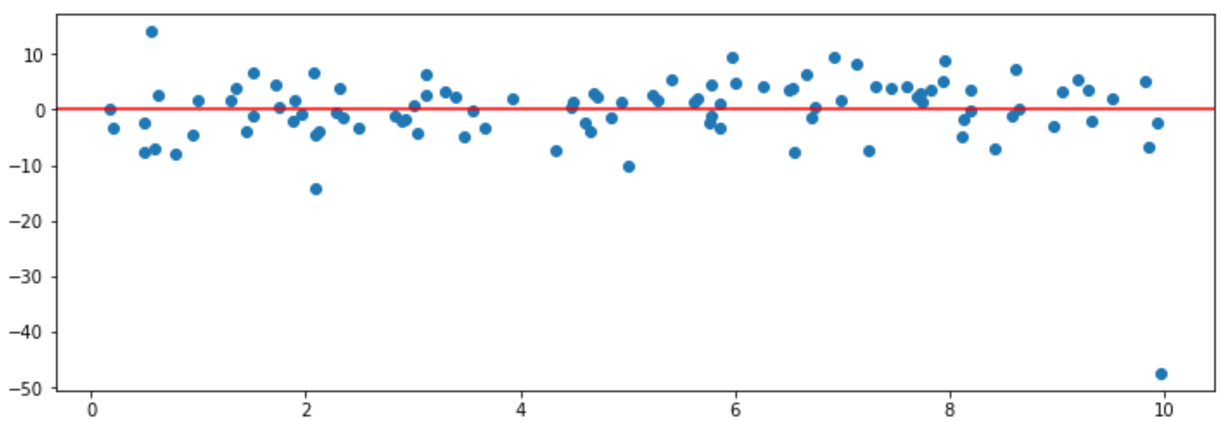


The plot above shows an apparent upward linear trend, where the remaining variability in the data around the line is not too major relative to the strength of the relationship between x and y. The strengh of the linear fit is explained by the R-squared value 0.979. There seems to be 3 possible outliers but the one at X=10 seems to be an influential point. It pulls the least squares line down on the right. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

In [23]:
```
#Histogram of Residuals
residual5=(df5.y-result5.predict(df5.x))
plt.hist(residual5, bins=10,color='c')
plt.show()
```



In [24]:
```
#Plot Residuals vs predictor
plt.figure(figsize=(12,4))
plt.plot(df5.x, residual5,'o')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
```



a. Linearity: As mentioned above, the scatter & line plot above confirms linearity of the dataset with a R-squared value of 0.806.

b. Nearly Normal Residuals: The residuals are nearly normal but has a left tail due to the outlier.

c. Constant variability: The variability of the data around the line remains constant with larger values of x, except the influential point.

d. Independent Observations: We assume observations are independent. (We don't have additional information of the data determine whether or not observations are independent.)

Set 6

In [25]:

```
#Read Set 6 worksheet from Module 6
df6=pd.read_excel(mod6, 'Set 6')
df6
```

Out[25]:

	y	x
0	25.8447	3.6921
1	27.4407	3.9201
2	55.8250	7.9750
...	...	...
98	32.1370	4.5910
99	18.3295	2.6185
100	250.4838	35.7834

101 rows × 2 columns

In [26]:
```
result6=sm.ols('df6.y~df6.x', data=df6).fit()
result6.summary()
```

Out[26]:

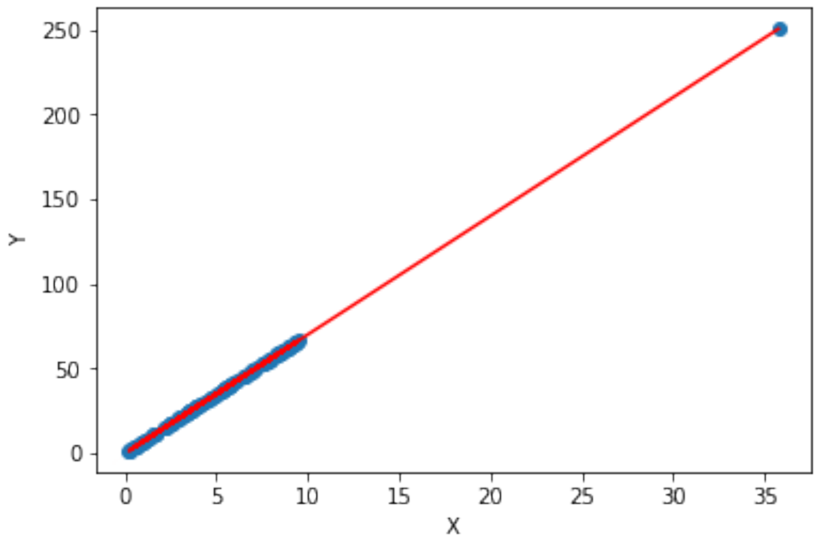
OLS Regression Results

Dep. Variable:	df6.y	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	3.534e+32
Date:	Tue, 03 Jul 2018	**Prob (F-statistic):**	0.00
Time:	14:27:24	**Log-Likelihood:**	3069.6
No. Observations:	101	**AIC:**	-6135.
Df Residuals:	99	**BIC:**	-6130.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.474e-14	2.48e-15	9.990	0.000	1.98e-14	2.96e-14
df6.x	7.0000	3.72e-16	1.88e+16	0.000	7.000	7.000

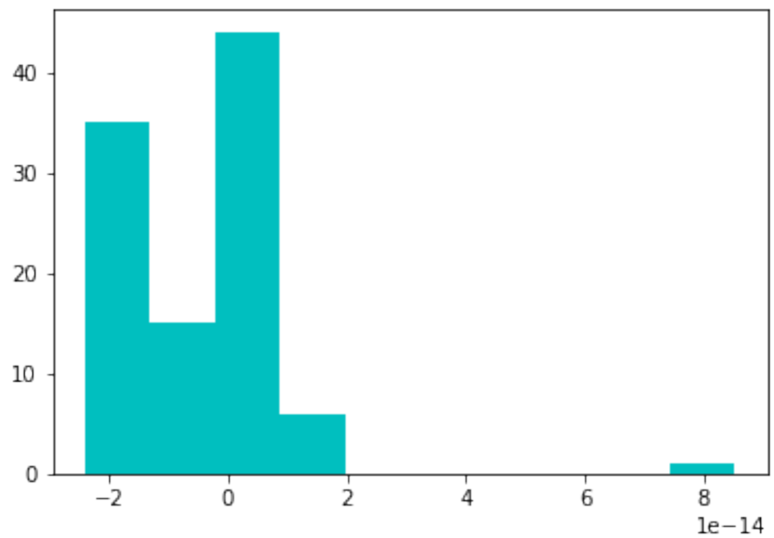
Omnibus:	76.924	**Durbin-Watson:**	1.631
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	750.604
Skew:	2.278	**Prob(JB):**	1.02e-163
Kurtosis:	15.554	**Cond. No.**	10.9

In [27]:
```
#Plot Y vs X (scatter plot) and a linear fitted line
plt.plot(df6.x, df6.y, 'o')
intercept, slope = result6.params
plt.plot(df6.x, intercept + slope * df6.x , 'r-', label='Fitted Line')
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```

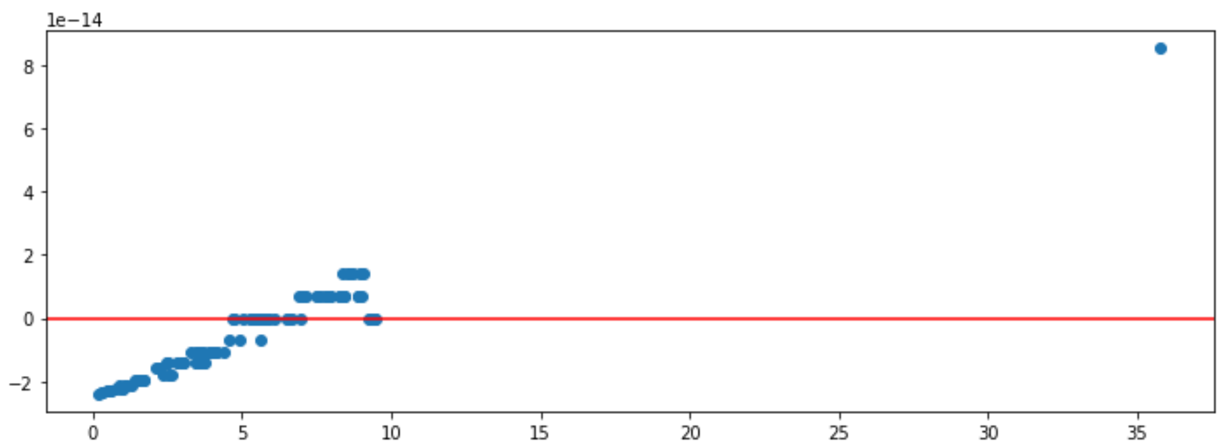


The plot above shows a very strong upward linear trend (R-squared: 1.000), where the points fall perfectly on the linear line. There is one outlier far right from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential.

In [28]:
```
#Histogram of Residuals
residual6=(df6.y-result6.predict(df6.x))
plt.hist(residual6, bins=10,color='c')
plt.show()
```



In [29]:
```
#Plot Residuals vs predictor
plt.figure(figsize=(12,4))
plt.plot(df6.x, residual6,'o')
plt.axhline(y=0, color='r', linestyle='-')
plt.show()
```



a. Linearity: As mentioned above, the dataset shows a strong linear relationship.

b. Nearly Normal Residuals: The residuals do not show a normal distribution: bimodal.

c. Constant variability: The variability of the data around the line does not remain constant with larger values of x but the residuals are quite close to the line.

d. Independent Observations: We assume observations are independent. (We don't have additional information of the data determine whether or not observations are independent.)