Regression Analysis in Statsmodels - Lab

Introduction

In the previous code along, we looked all the requirements for running an ols simple regression using statsmodels. We worked with a toy example to understand the process and all the necessary steps that must be performed. In this lab , we shall look at a slightly more complex example to study the impact of spendings in different advertising channels of total sales.

Objectives ¶

You will be able to:

- Set up an analytical question to be answered by regression analysis
- Study regression assumptions for real world datasets
- Visualize the results of regression analysis

Let's get started

In this lab, we will work with the "Advertising Dataset" which is a very popular dataset for studying simple regression. The dataset is available at Kaggle, but we have already downloaded for you. It is available as "Advertising.csv". We shall use this dataset to ask ourselves a simple analytical question:

The Question

Which advertising channel has a strong relationship with sales volume, and can be used to model and predict the sales.

Step 1: Read the dataset and inspect its columns and 5-point statistics

In [35]:
```
# Load necessary libraries and import the data
import pandas as pd
import matplotlib.pyplot as plt
plt.style.use('ggplot')
data = pd.read_csv('Advertising.csv', index_col=0)
```

In [36]:
```
# Check the columns and first few rows
data.head()
```

Out[36]:

	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

In [37]:
```
# Get the 5-point statistics for data
data.describe()
```

Out[37]:

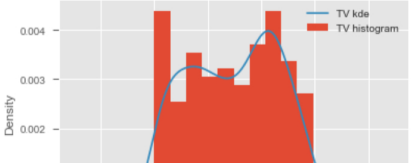
	TV	radio	newspaper	sales
count	200.000000	200.000000	200.000000	200.000000
mean	147.042500	23.264000	30.554000	14.022500
std	85.854236	14.846809	21.778621	5.217457
min	0.700000	0.000000	0.300000	1.600000
25%	74.375000	9.975000	12.750000	10.375000
50%	149.750000	22.900000	25.750000	12.900000
75%	218.825000	36.525000	45.100000	17.400000
max	296.400000	49.600000	114.000000	27.000000

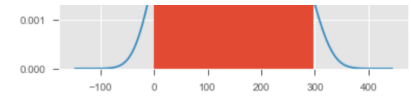
In [42]:
```
# Describe the contents of this dataset
```

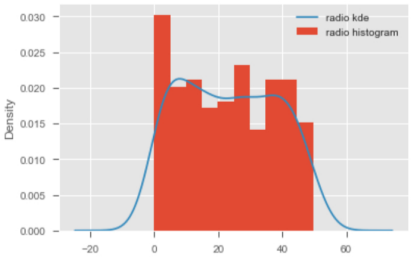
```
In every record, we have three predictors showing the advertising budget spent on TV, newspaper and radio and a target v
ariable (sales). The target variable shows the sales figure for each marketing campaign along with money spent on all th
ree channels.

Looking at means for predictors, most budget is spent on TV marketing , and least on radio.
```

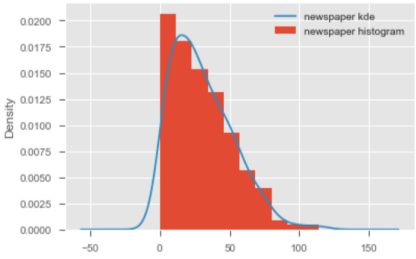
Step 2: Plot histograms with kde overlay to check for the normality of the predictors

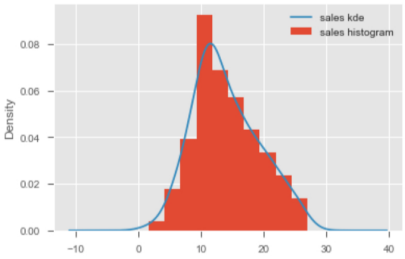
In [64]:
```
# For all the variables, check if they hold normality assumption
for column in data:
    data[column].plot.hist(normed=True, label = column+' histogram')
    data[column].plot.kde(label =column+' kde')
    plt.legend()
    plt.show()
```











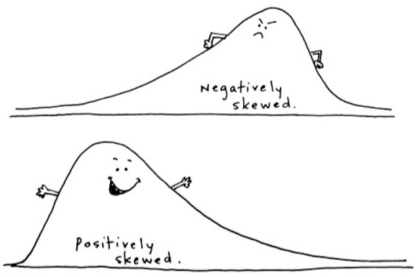
In [ ]:
```
# Record your observations on normality here
```

No variable is "perfectly" normal, but these do tend to follow an overall normal pattern. We see major skew in the newsp aper predictor which could be problematic towards analysis.

TV and radio are still pretty symmetrical distributions and can be used as predictors

The target variable "sales" is normally distributed with just a gentle skew

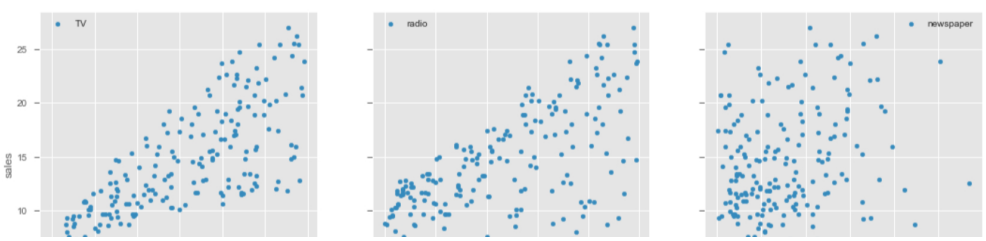
Remember . Nothing is perfect . So be positive

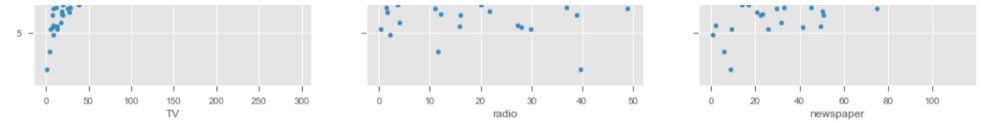


Step 3: Test for the linearity assumption

Use scatterplots to plot each predictor against the target variable

In [57]:
```
# visualize the relationship between the preditors and the target using scatterplots
fig, axs = plt.subplots(1, 3, sharey=True, figsize=(18, 6))
for idx, channel in enumerate(['TV', 'radio', 'newspaper']):
    data.plot(kind='scatter', x=channel, y='sales', ax=axs[idx], label=channel)
plt.legend()
plt.show()
```





In [ ]:
```
# Record yor observations on linearity here
```

```
TV seems to be a good feature due to highly linear relationship with sales
radio shows a linear pattern as well but there a higher level of variance in there than TV
newspaper is worse, there is too much variance along y-axis and theres no clear linear relationship between newspaper an
d sales
```

Conclusion so far !

Based on above initial checks, we can confidently say that TV and radio appear to be good predictors for our regression analysis. Newspaper is very heavily skewed and also doesnt show any clear linear relationship with the target.

> We shall move ahead with our analysis using TV and radio , and count out the newspaper due to the fact that data violates ols assumptions

Note: Kurtosis can be dealt with using techniques like log normalization to "push" the peak towards the center of distribution. We shall talk about this in the next section.

Step 4: Run a simple regression in `statsmodels` with TV as a predictor

In [83]:
```
# import libraries
import statsmodels.api as sm
import statsmodels.formula.api as smf

# build the formula
f = 'sales~TV'
# create a fitted model in one line
model = smf.ols(formula=f, data=data).fit()
```

Step 5: Get regression diagnostics summary

In [ ]:
```
model.summary()
```

Record your observations on "Goodness of fit"

```
R-squared value is 0.61 i.e. 61% of variance in the target variable can be explained using the spendings on TV.

The Intercept: A "unit" increase in TV spending is associated with a 0.0475 "unit" increase in Sales. OR An additional
 1,000 spent on TV is associated with an increase in sales of 47.5
```

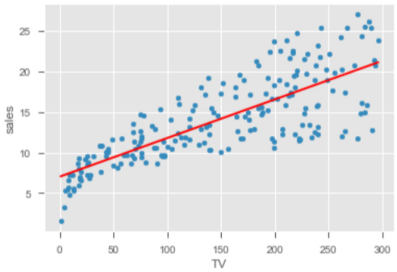
Note here that the coefficients represent associations, not causations

Step 6: Draw a prediction line with data points omn a scatter plot for X (TV) and Y (Sales)

Hint: We can use `model.predict()` functions to predict the start and end point of of regression line for the minimum and maximum values in the 'TV' variable.

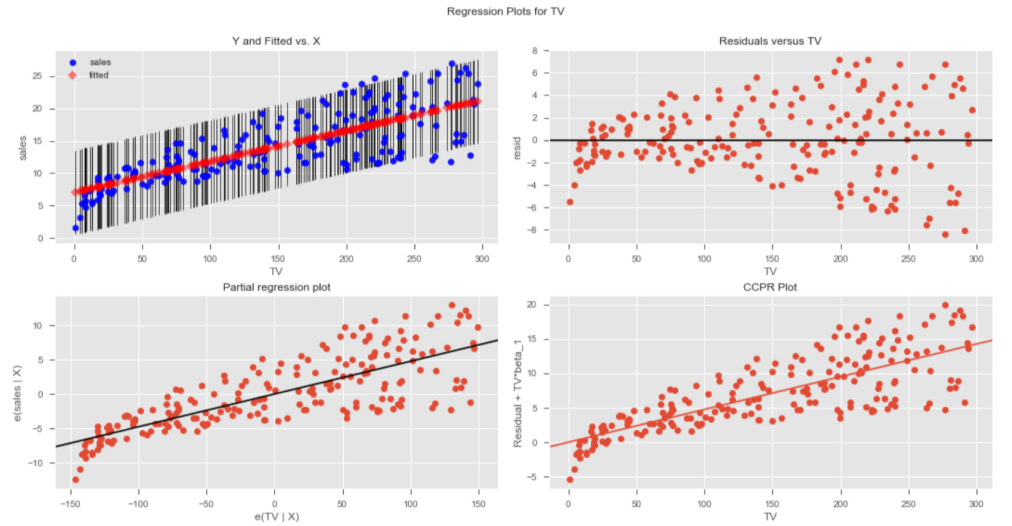
In [85]:
```
# create a DataFrame with the minimum and maximum values of TV
X_new = pd.DataFrame({'TV': [data.TV.min(), data.TV.max()]})
print(X_new.head())

# make predictions for those x values and store them
preds = model.predict(X_new)
print (preds)

# first, plot the observed data and the least squares line
data.plot(kind='scatter', x='TV', y='sales')
plt.plot(X_new, preds, c='red', linewidth=2)
plt.show()
```

```
      TV
0    0.7
1  296.4
0     7.065869
1    21.122454
dtype: float64
```



Step 7: Visualize the error term for variance and heteroscedasticity

In [87]:
```
fig = plt.figure(figsize=(15,8))
fig = sm.graphics.plot_regress_exog(model, "TV", fig=fig)
plt.show()
```



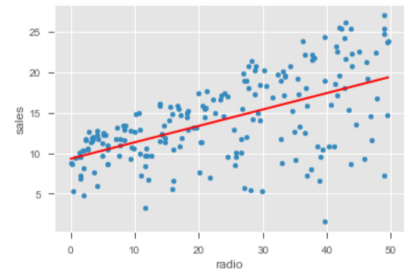
In [ ]:
```
# Record Your observations on residuals
```

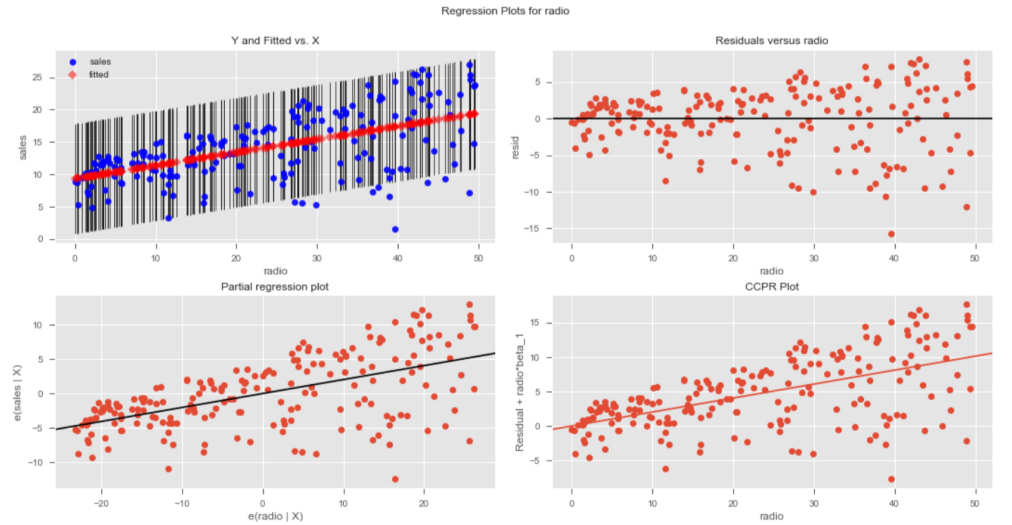
From the first and second plot in the first row, we see that the variance is creating a cone-shape which is a sign of he teroscedasticity. i.e. the residuals are not normally distributed . This breaks the assumption.

Next, repeat above for radio and go through the same process, recording your observations

In [89]:
```
f = 'sales~radio'
model = smf.ols(formula=f, data=data).fit()
print ('R-Squared:',model.rsquared)
print (model.params)
X_new = pd.DataFrame({'radio': [data.radio.min(), data.radio.max()]});
preds = model.predict(X_new)
data.plot(kind='scatter', x='radio', y='sales');
plt.plot(X_new, preds, c='red', linewidth=2);
plt.show()
fig = plt.figure(figsize=(15,8))
fig = sm.graphics.plot_regress_exog(model, "radio", fig=fig)
plt.show()
```

```
R-Squared: 0.33203245544529525
Intercept    9.311638
radio        0.202496
dtype: float64
```





In [77]:
```
model.summary()
```

Out[77]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.332

Model:	OLS	**Adj. R-squared:**	0.329
Method:	Least Squares	**F-statistic:**	98.42
Date:	Fri, 12 Oct 2018	**Prob (F-statistic):**	4.35e-19
Time:	20:52:55	**Log-Likelihood:**	-573.34
No. Observations:	200	**AIC:**	1151.
Df Residuals:	198	**BIC:**	1157.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.3116	0.563	16.542	0.000	8.202	10.422
radio	0.2025	0.020	9.921	0.000	0.162	0.243

Omnibus:	19.358	**Durbin-Watson:**	1.946
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21.910
Skew:	-0.764	**Prob(JB):**	1.75e-05
Kurtosis:	3.544	**Cond. No.**	51.4

In [79]:
```
# Record your observations here for goodnes of fit
```

```
As a predictor, radio performs worse than TV.

It has higher amount of skewness and kurtosis than TV

A very low R_squared explaining only 33% of variance in the target variable

A "unit" increase in radio spending is associated with a 0.2025 "unit" increase in Sales. OR An additional 1,000 spent o
n TV is associated with an increase in sales of 20.02

There is obvious heteroscedasticity as with the case of TV
```

The Answer

Based on above analysis, we can conclude that none of the two chosen predictors is ideal for modeling a relationship with the sales volumes. `Newspaper` clearly violated normality and linearity assumptions. TV and radio did not provide a high value for co-efficient of determination - TV performed slightly better than the radio. There is obvious heteroscdasticity in the residuals for both variables.

> We can either look for further data, perform extra pre-processing or use more advanced techniques.

Remember there are lot of technqiues we can employ to FIX this data.

Whether we should call TV the "best predictor" or label all of them "equally useless", is a domain specific question and a marketing manager would have a better opinion on how to move forward with this situation.

In the following lesson, we shall look at the more details on interpreting the regression diagnostics and confidence in the model.

Summary

In this lesson, we ran a complete regression analysis with a simple dataset. We looked for the regression assumptions pre and post the analysis phase. We also created some visualizations to develop a confidence on the model and check for its goodness of fit.