Here are some recommended packages, not all are required and depends on your solution.

In [105]:

```
# imports
import pandas as pd
import seaborn as sns
import statsmodels.formula.api as smf
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.sandbox.regression.predstd import wls_prediction_std

# allow plots to appear directly in the notebook
%matplotlib inline
```

Questions

You are a consultant for a company that sells widgets. They have historical data on their sales on their investments in advertising in various media outlets, including TV, radio, and newspapers. On the basis of this data, how should they be spending their advertising money in the future?

Your analysis should answer the following questions:

Is there a relationship between ads and sales?

How strong is that relationship?

Which ad types contribute to sales?

What is the effect of each ad type of sales?

Given ad spending in a particular market, can sales be predicted?

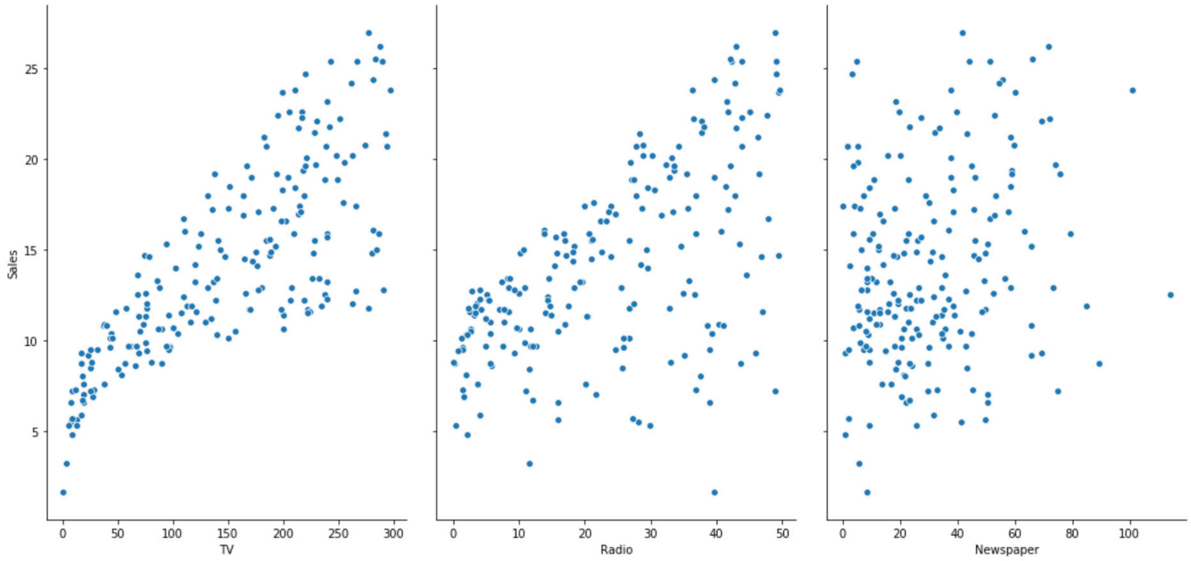
In [102]:

```
# read data into a DataFrame, this is money spent on different medias
data = pd.read_csv('https://raw.githubusercontent.com/lneisenman/isl/master/data/Advertising.csv', index_col=0)
print(data.head())
# visualize the relationship between the features and the response using scatterplots
sns.pairplot(data, x_vars=['TV','Radio','Newspaper'], y_vars='Sales', height=7, aspect=0.7)
```

```
      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
```

Out[102]:

```
<seaborn.axisgrid.PairGrid at 0x15243a9d710>
```



In the lecture, we covered how to perform a linear regression model. We did not however explore how "good" this model is. The task below will have you identifying ways to evaluate a linear regression model.

Machine learning focuses on what the model predicts. If you would like to dive into the meaning of fit parameters within the model, other tools are available, including the Statsmodels Python package. Take some time to look at this package (https://www.statsmodels.org/stable/regression.html) and also an example of evaluating a linear regression (https://www.statsmodels.org/stable/examples/notebooks/generated/gls.html).

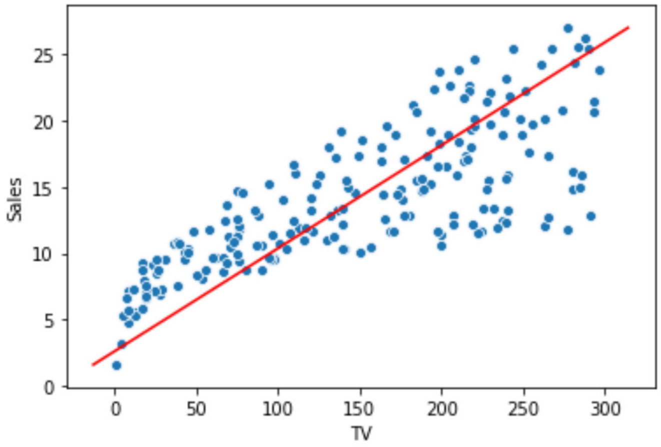
Similar to Scikit-learn, one can calculate the intercept and coefficient for a linear fit for a set of data.

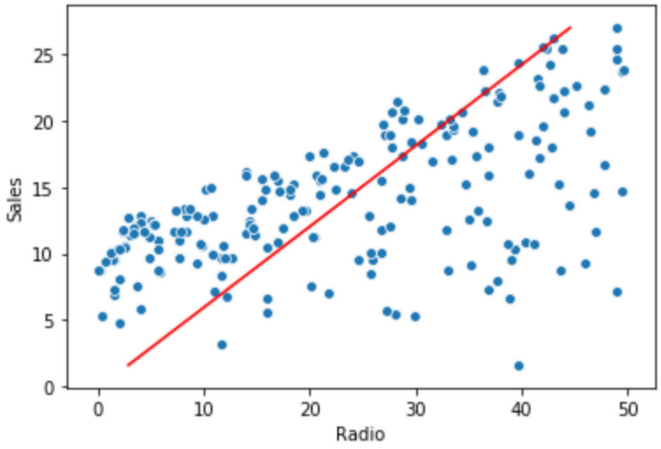
In [103]:

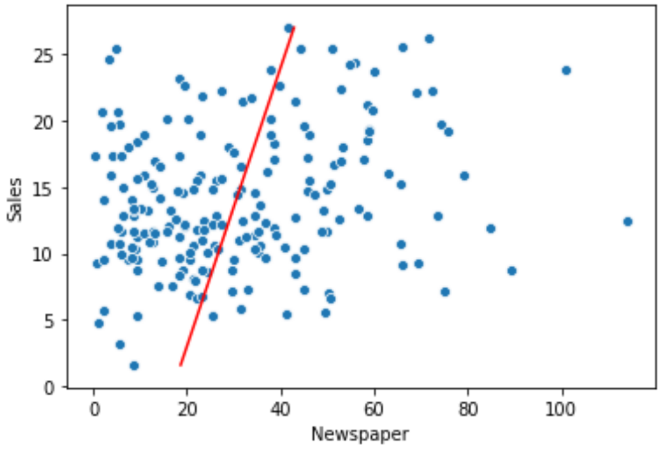
```
#olsTV=smf.ols(formula='TV ~ Sales', data=data).fit()
print(data['TV'].shape)
olsTV=smf.ols(formula='TV ~ Sales', data=data).fit()
print(olsTV.params)

olsRadio = smf.ols(formula='Radio ~ Sales', data=data).fit()
print(olsRadio.params)

olsPaper = smf.ols(formula='Newspaper ~ Sales', data=data).fit()
print(olsPaper.params)

data['TVfit']=olsTV.fittedvalues
data['RadioFit']=olsRadio.fittedvalues
data['PaperFit']=olsPaper.fittedvalues

fig, axs1 = plt.subplots()
sns.lineplot(x='TVfit',y='Sales',data=data, ax=axs1, color='r')
sns.scatterplot(x='TV',y='Sales',data=data, ax=axs1)

fig2,axs2 = plt.subplots()
sns.lineplot(x='RadioFit',y='Sales',data=data, ax=axs2, color='r')
sns.scatterplot(x='Radio',y='Sales',data=data, ax=axs2)

fig3,axs3 = plt.subplots()
sns.lineplot(x='PaperFit',y='Sales',data=data, ax=axs3, color='r')
sns.scatterplot(x='Newspaper',y='Sales',data=data, ax=axs3)
```

```
(200,)
Intercept   -33.450228
Sales        12.871651
dtype: float64
Intercept     0.271298
Sales         1.639701
dtype: float64
Intercept    17.191090
Sales         0.952962
dtype: float64
```

```
Out[103]: <matplotlib.axes._subplots.AxesSubplot at 0x1524420eac8>
```





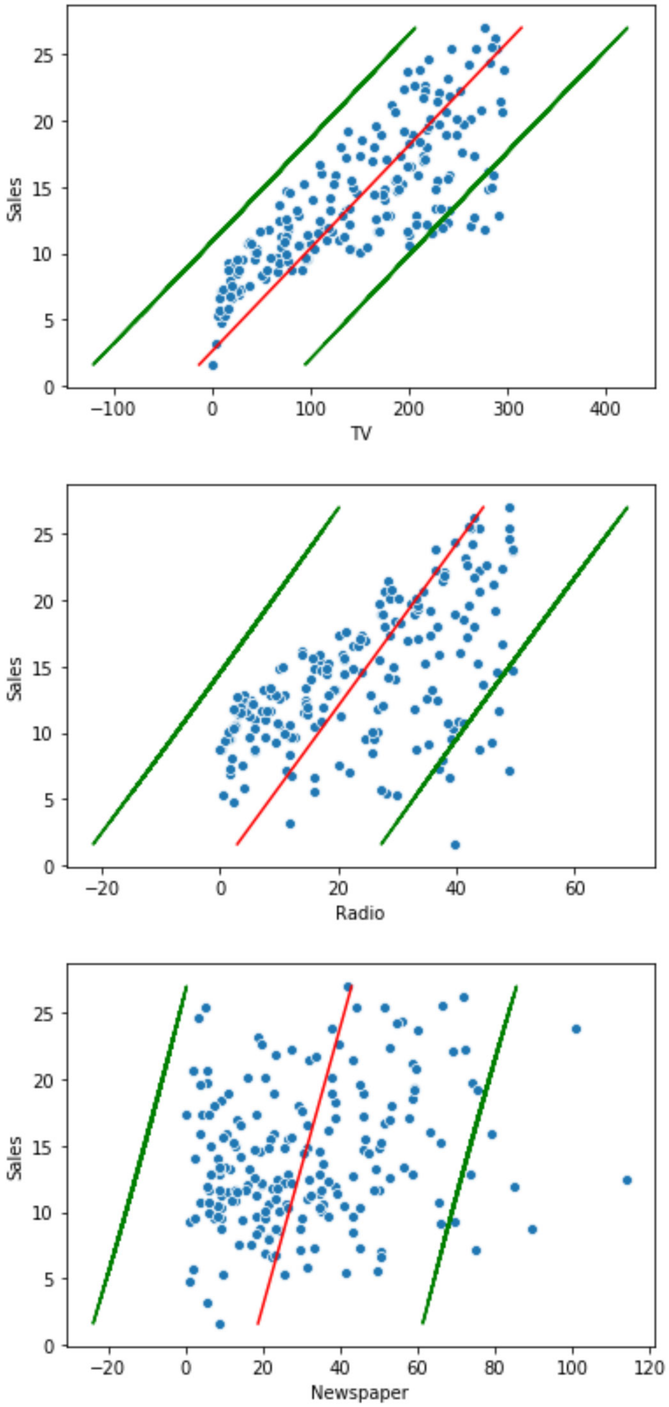


A confidence interval can be used to describe a linear model. How would you calculate the confidence interval of this model and what does this confidence interval mean?

In [120]:

```
#Make graphs of confidence intervals
yVal=data['Sales'].values

prstd, iv_lTV, iv_uTV = wls_prediction_std(olsTV)
fig, axs2_1 = plt.subplots()
sns.lineplot(x='TVfit',y='Sales',data=data, ax=axs2_1, color='r')
sns.scatterplot(x='TV',y='Sales',data=data, ax=axs2_1)
axs2_1.plot(iv_uTV, yVal, 'g')
axs2_1.plot(iv_lTV, yVal, 'g')

prstd, iv_lR, iv_uR = wls_prediction_std(olsRadio)
fig2, axs2_2 = plt.subplots()
sns.lineplot(x='RadioFit',y='Sales',data=data, ax=axs2_2, color='r')
sns.scatterplot(x='Radio',y='Sales',data=data, ax=axs2_2)
axs2_2.plot(iv_uR, yVal, 'g')
axs2_2.plot(iv_lR, yVal, 'g')

prstd, iv_lP, iv_uP = wls_prediction_std(olsPaper)
fig3, axs2_3 = plt.subplots()
sns.lineplot(x='PaperFit',y='Sales',data=data, ax=axs2_3, color='r')
sns.scatterplot(x='Newspaper',y='Sales',data=data, ax=axs2_3)
axs2_3.plot(iv_uP, yVal, 'g')
axs2_3.plot(iv_lP, yVal, 'g')
```

```
Out[120]: [<matplotlib.lines.Line2D at 0x1524590e7b8>]
```



Confidence interval can be calculated using the wls_prediction_std from the statsmodel api. Mathematically, 95% confidence interval is the x_estimate +- 1.96*(standard error), 1.96 is the critical value for a CI of 95%. Confidence interval gives us a range where we are like to find the true value in. With an alpha of 0.05, confidence interval shows where 95% of values fall within 1.96x the standard deviation of the model.

Other metrics that are used to describe the appropriateness of a model is a p-value. How would you calculate the p-value and r-squared values of the model? What do these values mean?

```
In [82]: olsTV.summary()
```

```
Out[82]:
OLS Regression Results

     Dep. Variable:                TV         R-squared:     0.612
             Model:               OLS     Adj. R-squared:     0.610
            Method:     Least Squares        F-statistic:     312.1
              Date:  Sun, 06 Oct 2019  Prob (F-statistic):  1.47e-42
              Time:          20:45:58     Log-Likelihood:   -1079.2
 No. Observations:               200                AIC:     2162.
      Df Residuals:               198                BIC:     2169.
          Df Model:                 1
   Covariance Type:         nonrobust

                coef  std err       t  P>|t|  [0.025   0.975]
 Intercept  -33.4502   10.897  -3.070  0.002  -54.939  -11.961
     Sales   12.8717    0.729  17.668  0.000   11.435   14.308

       Omnibus:  21.952    Durbin-Watson:     1.973
 Prob(Omnibus):   0.000  Jarque-Bera (JB):   26.224
          Skew:   0.882          Prob(JB):  2.02e-06
      Kurtosis:   3.193          Cond. No.      43.2


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

```
In [83]: olsRadio.summary()
```

```
Out[83]:
OLS Regression Results

    Dep. Variable:             Radio         R-squared:      0.332
           Model:                OLS    Adj. R-squared:      0.329
          Method:      Least Squares       F-statistic:      98.42
            Date:   Sun, 06 Oct 2019  Prob (F-statistic): 4.35e-19
            Time:           20:46:03    Log-Likelihood:    -782.49
No. Observations:                200               AIC:      1569.
    Df Residuals:                198               BIC:      1576.
        Df Model:                  1
 Covariance Type:          nonrobust

             coef  std err      t  P>|t|  [0.025  0.975]
Intercept  0.2713    2.472  0.110  0.913  -4.604   5.146
    Sales  1.6397    0.165  9.921  0.000   1.314   1.966

      Omnibus: 15.769    Durbin-Watson:    1.980
Prob(Omnibus):  0.000 Jarque-Bera (JB):   17.838
         Skew:  0.732         Prob(JB): 0.000134
     Kurtosis:  2.991         Cond. No.     43.2


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

```
In [84]: olsPaper.summary()
```

```
Out[84]:
OLS Regression Results

      Dep. Variable:        Newspaper        R-squared:    0.052
             Model:               OLS    Adj. R-squared:    0.047
            Method:     Least Squares       F-statistic:    10.89
              Date:  Sun, 06 Oct 2019 Prob (F-statistic): 0.00115
              Time:          20:46:06    Log-Likelihood:  -894.12
  No. Observations:               200               AIC:    1792.
      Df Residuals:               198               BIC:    1799.
          Df Model:                 1
   Covariance Type:         nonrobust

               coef  std err      t  P>|t|  [0.025  0.975]
  Intercept 17.1911    4.320  3.980  0.000   8.672  25.710
      Sales  0.9530    0.289  3.300  0.001   0.383   1.523

        Omnibus: 24.387      Durbin-Watson:    1.882
  Prob(Omnibus):  0.000   Jarque-Bera (JB):   29.955
           Skew:  0.834           Prob(JB): 3.13e-07
       Kurtosis:  3.902           Cond. No.     43.2


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

r-squared values are calculated by taking the (sum of squared total error - sum of squared residual error)/sum of squared total error. r-squared value represents ratio of the total variance in the data explained by the model. The closer the r-squared value is to 1 the better the fit.

p-value in regression shows is there is indeed a relationship between the x and the y. The smaller the p-value the lower the probability that x and y are not related. Having a P<0.05 means that there is a significant relationship between x and y. p-values can be calculated using the t-test.

Discussion:

There appears to be a strong relationship between TV ads and sales, a medium relationship between radio ad and sales and a weak relationship between newspaper ad and sales.

From r-squared value, 61% of the variance can be explained by TV sales making it the strongest contributor to sales followed by radio at 33% and newspaper at 5%. Newpaper had a very low r-square value, newspaper linear model does not fit the data well.

Based on the p-test, all three ad category were statistically significant at alpha of 0.05. In combination with the r-squared value, newspaper might have significant effect on sales as there does seem to be a cluster on the bottom left of the plot, a linear model might just not be the best model to describe it's effect. Linear model for tv ads is a good fit and is significant meaning it is a good model that can could potentially be used for prediction. Radio has a low r-squared but is significant so another model could still be explored for a better fit.

Base on coeffecient of the linear regression model. TV ads have a large effect on sales with a slope of 12 followed by radio with an effect of 1.6 and newspaper with an coefficient of less than 1. This suggest that TV ads contributes the most to sales.

Based on data, sales could be predicted based on TV ads to a certain extent. However, there are many other factors besides the ad medium such as quality of add and the general state of the economy that would affect sales.

In [ ]: