Linear Regression

Vahid Partovi Nia

Lecture 02: Simple and Multiple Linear Regression

30 October 2018

Outline

1. Terminology
2. Advertisement
3. Simple Linear Regression
4. Multiple Linear Regression
5. Education

Equivalent terminologies

Terminology

Advertisement

Simple Linear Regression

Multiple Linear Regression

Education

- y: dependent variable, response variable, output variable
- x: independent variable, explanatory variable, input variable, feature.

Terminology

Advertisement

Simple Linear Regression

Multiple Linear Regression

Education

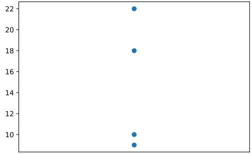
- Regression: y is continuous
- Classification: y is discrete

Sales prediction

Terminology
Advertisement
Simple Linear Regression
Multiple Linear Regression
Education

$$y_1 = 22, \quad y_2 = 10, \quad y_3 = 9, \quad y_4 = 18$$
$$y_i = \beta_0 + \varepsilon_i$$

- What is $\hat{y}_i$?
- What is $\hat{\beta}_0$?
- Least squares: $\min \frac{1}{5}\{(22-\beta_0)^2 + \cdots + (18-\beta_0)^2$
- $\hat{\beta}_0 = \frac{1}{5}(22 + \cdots + 18)$



Advertisement

Suppose we have a fixed budget of advertisement to increase sales.

Problem:

How do you distribute advertisement budget between different advertisement methods?

Advertisement

Suppose we have a fixed budget of advertisement to increase sales.

Problem:

How do you distribute advertisement budget between different advertisement methods?

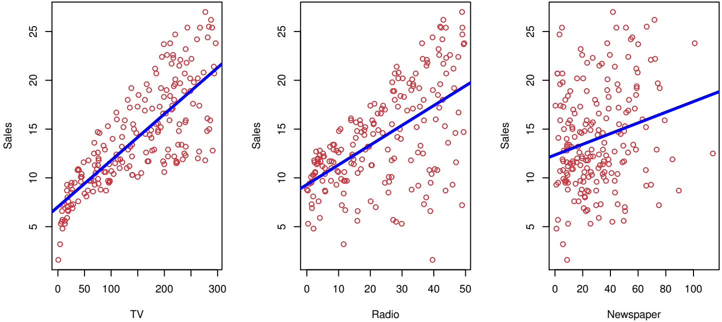
TV, Radio, Newspaper, Online, etc.

Question:

- Does advertisement affect sale?
- How do we predict sale?
- What is y what is x?
- Is it a regression or a classification?

Terminology
Advertisement
Simple Linear Regression
Multiple Linear Regression
Education

Sales: y



TV: x_1 Radio: x_2 Newspaper: x_3

$$\text{Learning: Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper}) + \varepsilon$$

Sale prediction simplification

$$
\begin{aligned}
\text{Sales} &\approx f(\text{TV}, \text{Radio}, \text{Newspaper}) \\
&\Downarrow \\
\text{Sales} &\approx f_1(\text{TV}) + f_2(\text{Radio}) + f_3(\text{Newspaper}) \\
&\Downarrow \\
\text{Sales} &\approx f_1(\text{TV}) \\
&\Downarrow \\
y &\approx \beta_0 + \beta_1 \text{TV} \\
&\Downarrow \\
y &\approx \beta_0
\end{aligned}
$$

Python

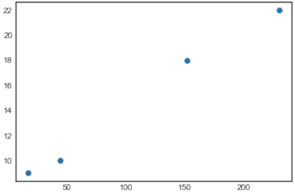
Step 1

- Load “Advertising.csv”
  - path='/Users/Desktop/datafiles/'
  - filename=path+'Advertising.csv'
- `import pandas as pd`
- `import numpy as np`
- Take the mean of “sales”

```
import pandas as pd
path='data/'
filename = path+'Advertising.csv'
advertising = pd.read_csv(filename)
```

```
import numpy as np
np.mean(advertising['sales'])
```

$$y_1 = 22 \quad y_2 = 10 \quad y_3 = 9 \quad y_4 = 18$$
$$x_{11} = 230 \quad x_{12} = 44 \quad x_{13} = 17 \quad x_{14} = 151$$



$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

What is $\hat{y}_i$?

Terminology
Advertisement
Simple Linear Regression
Multiple Linear Regression
Education

Step 2: Predict Sales using TV

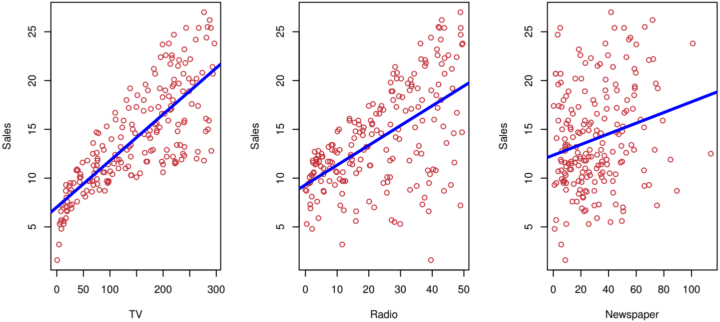
- Load "LinearRegression" from sklearn
- Initialize the model
- Feed the data
- Scatter plot Sales versus TV
- Add the predicted line

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
```

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()

lr.fit(X = advertising[ ['TV'] ], y = advertising['sales'])
print(lr.intercept_, lr.coef_)
```

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()

lr.fit(X = advertising[ ['TV'] ], y = advertising['sales'])
print(lr.intercept_, lr.coef_)

import matplotlib.pyplot as plt
%matplotlib inline

plt.plot(advertising.TV, advertising.sales, 'or', mfc='none');
plt.plot(advertising.TV, lr.intercept_+lr.coef_*advertising.TV, '-b');

plt.xlabel('TV');
plt.ylabel('sales');
```

Combine models

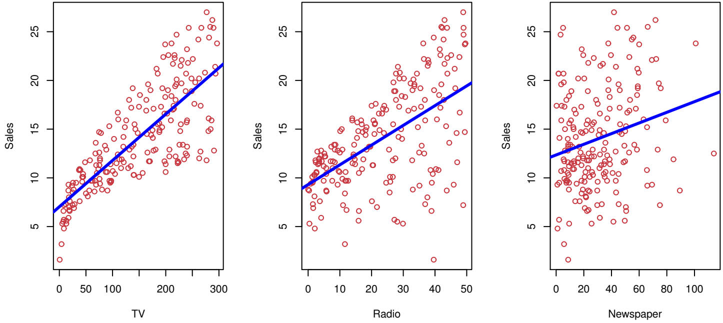


TV: x_1 Radio: x_2 Newspaper: x_3

$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \varepsilon$$

Combine models

Terminology
Advertisement
Simple Linear Regression
Multiple Linear Regression
Education



TV: x_1 Radio: x_2 Newspaper: x_3

$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \varepsilon$$

Predict for $\text{TV} = 250, \text{Radio} = 30, \text{Newspaper} = 20$

Terminology

Advertisement

Simple Linear Regression

Multiple Linear Regression

Education

```
lr = LinearRegression()
```

```
lr = LinearRegression()

lr.fit(X = advertising[ ['TV', 'radio', 'newspaper'] ],
       y = advertising['sales'])

x = np.array([250, 30, 20] )
lr.predict(x.reshape(1,3))
```

Terminology
Advertisement
Simple Linear Regression
Multiple Linear Regression
Education

```
lr = LinearRegression()

lr.fit(X = advertising[ ['TV', 'radio', 'newspaper'] ],
       y = advertising['sales'])

x = np.array([250, 30, 20] )
lr.predict(x.reshape(1,3))


x = np.array([250, 30, 20, 249, 29, 19] )
lr.predict(x.reshape(2,3))
```

```
import statsmodels.formula.api as smf
model = smf.ols(formula='sales ~ TV + radio + newspaper',
                data = advertising)
lr = model.fit()
lr.summary()
```

ols summary

Terminology

Advertisement

Simple Linear Regression

Multiple Linear Regression

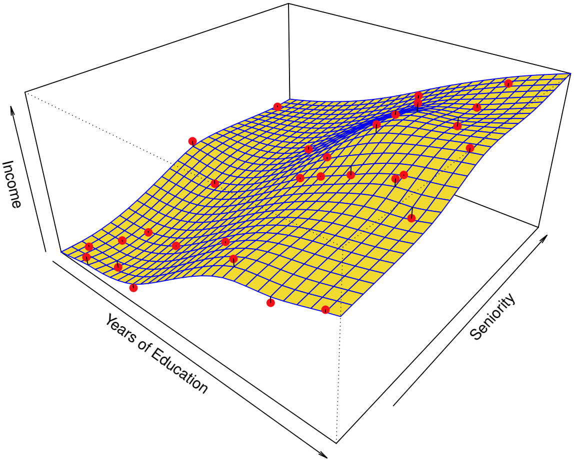
Education

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Sat, 31 Mar 2018	**Prob (F-statistic):**	1.58e-96
Time:	17:41:29	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	151.241
Skew:	-1.327	**Prob(JB):**	1.44e-33
Kurtosis:	6.332	**Cond. No.**	454.

Income: y



Years of Education: x_1 Seniority: x_2

$y \approx f(x_1, x_2)$

Terminology

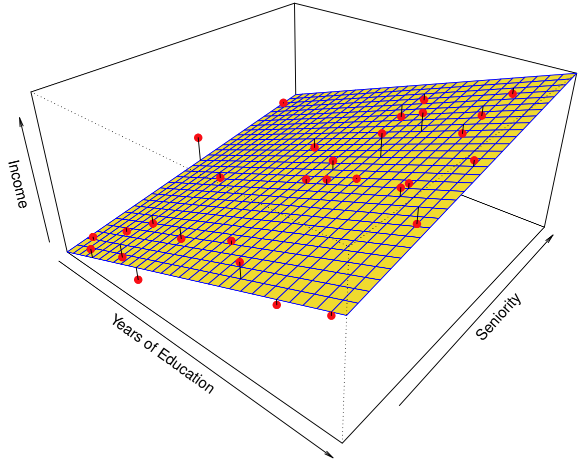
Advertisement

Simple Linear Regression

Multiple Linear Regression

Education

Income: y



$$y \approx f_1(x_1) + f_2(x_2)$$

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$