IIT ROORKEE

NPTEL ONLINE
CERTIFICATION COURSE

REGRESSION

Linear Regression

Dr. Ramesh Anbanandam

DEPARTMENT of Management Studies

Simple Linear Regression

- Simple Linear Regression Model
- Least Squares Method
- Coefficient of Determination
- Model Assumptions
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction

Empirical Models

- Many problems in engineering and science involve exploring the relationships between two or more variables
- Regression analysis is a statistical technique that is very useful for these types of problems
- This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes

Empirical Models Example

- As an illustration, consider the data in the table.
- In this table y is the purity of oxygen produced in a chemical distillation process, and x is the percentage of hydrocarbons that are present in the main condenser of the distillation unit.

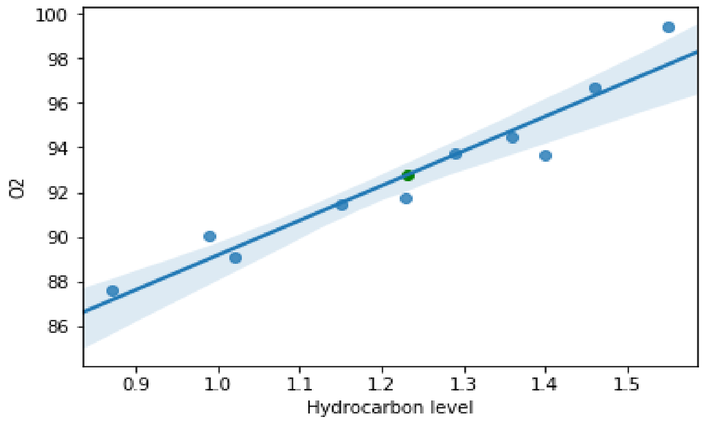
Hydrocarbon level (X)	Purity (Y)
0.99	90.01
1.02	89.05
1.15	91.43
1.29	93.74
1.46	96.73
1.36	94.45
0.87	87.59
1.23	91.77
1.55	99.42
1.4	93.65

Using python for plotting the data

```
In [20]: data = pd.read_excel('C:/Users/Somi/Desktop/reg2.xlsx')
```

```
In [19]: x= data['Hydrocarbon level']
         y = data['O2']
         plt.figure()
         sns.regplot(x,y,fit_reg= True)
         plt.scatter(np.mean(x), np.mean(y), color = "green")
```

```
Out[19]: <matplotlib.collections.PathCollection at 0x21ada0ab1d0>
```



Simple Linear Regression Model

- The equation that describes how *y* is related to *x* and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

β_0 and β_1 are called parameters of the model,

ε is a random variable called the error term.

Simple Linear Regression Equation

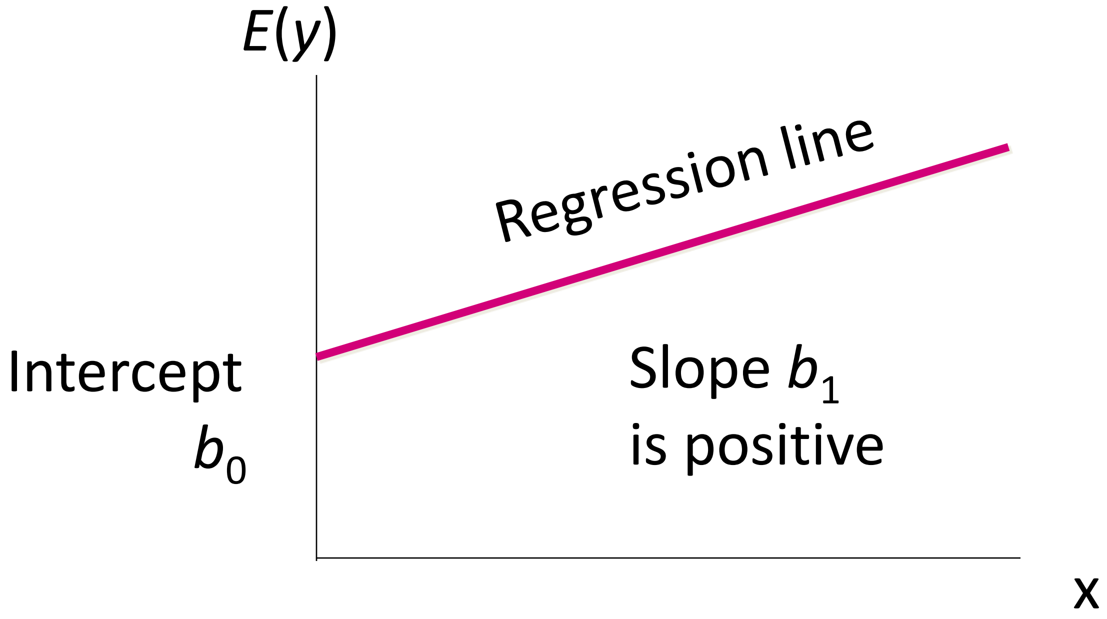
The simple linear regression equation is:

$$E(y) = \beta_0 + \beta_1 x$$

- Graph of the regression equation is a straight line.
- β_0 is the *y* intercept of the regression line.
- $\beta 1$ is the slope of the regression line.
- *E*(*y*) is the expected value of *y* for a given *x* value.

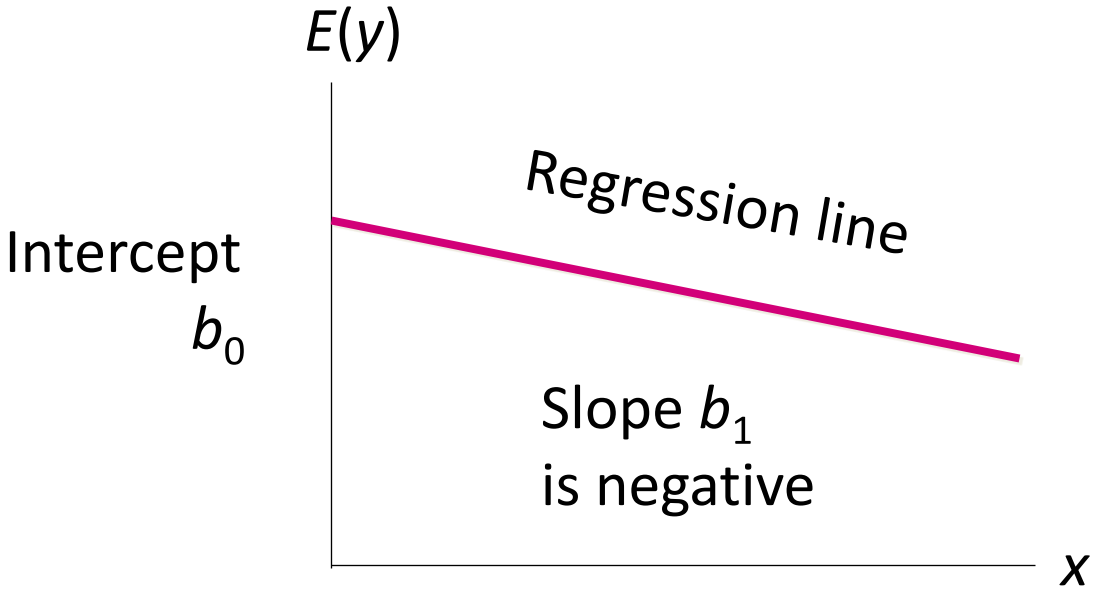
Simple Linear Regression Equation

Positive Linear Relationship



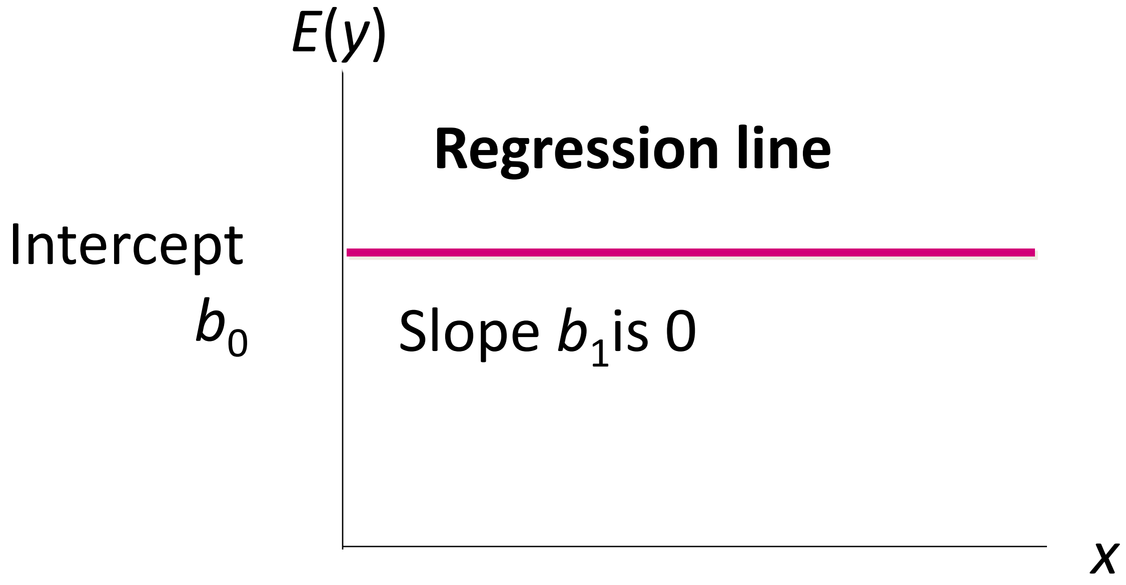
Simple Linear Regression Equation

Negative Linear Relationship



Simple Linear Regression Equation

No Relationship



Estimated Simple Linear Regression Equation

- The estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x$$

  - The graph is called the estimated regression line.
    - b_0 is the y intercept of the line.
    - b_1 is the slope of the line.
    - $\hat{y}$ is the estimated value of y for a given x value.

Least Squares Method

- Least Squares Criterion

$$\boxed{\min \sum (y_i - \hat{y}_i)^2}$$

where:

y_i = <u>observed</u> value of the dependent variable for the ith observation

$\hat{y}_i$ = <u>estimated</u> value of the dependent variable for the ith observation

Estimation Process

Regression Model

$y = \beta_0 + \beta_1 x + e$

Regression Equation

$E(y) = \beta_0 + \beta_1 x$

Unknown Parameters

β_0, β_1

Sample Data:

x	y
x_1	y_1
.	.
.	.
x_n	y_n

Estimated Regression Equation

$$\hat{y} = b_0 + b_1 x$$

Sample Statistics

b_0, b_1

b_0 **and** b_1 **provide estimates of** β_0 **and** β_1

$$\text{Squared Error (SE)} = \left(y_1 - (mx_1+b)\right)^2 + \left(y_2 - (mx_2+b)\right)^2 + \ldots \left(y_n - (mx_n+b)\right)^2$$

$$= y_1^2 - 2y_1(mx_1+b)+(mx_1+b)^2$$

$$+y_2^2 - 2y_2(mx_2+b)+(mx_2+b)^2$$

$$+\ldots$$

$$+y_n^2 - 2y_n(mx_n+b)+(mx_n+b)^2$$

$$= y_1^2 - 2x_1y_1m - 2y_1b+m^2x_1^2 + 2mx_1b+b^2$$

$$+y_2^2 - 2x_2y_2m - 2y_2b+m^2x_2^2 + 2mx_2b+b^2$$

$$+\ldots\ldots$$

$$+y_n^2 - 2x_ny_nm - 2y_nb+m^2x_n^2 +2mx_nb+b^2$$

$$
\begin{aligned}
&=(y_1^2 + y_2^2 + \ldots + y_n^2) \\
&-2m\,(x_1y_1 + x_2y_2 + \ldots + x_ny_n) \\
&-2b(y_1 + y_2 + \ldots. + y_n) \\
&+m^2(x_1^2 + x_2^2 + .. + x_n^2) \\
&+2mb(x_1 + x_2 + .. + x_n) \\
&+(b^2 + b^2 + \ldots + b^2)
\end{aligned}
$$

$$= n\,\overline{y^2} - 2mn\,\overline{x\,y} - 2bn\,\overline{y} + m^2 n\,\overline{x^2} + 2mbn\overline{x} + nb^2$$

$$\mathrm{SE} = n\overline{y^2} - 2mn\,\overline{x\,y} - 2bn\,\overline{y} + m^2 n\,\overline{x^2} + 2mbn\,\overline{x} + nb^2$$

$$\frac{\partial(\mathrm{SE})}{\partial m} = -2n\,\overline{x\,y} + 2mn\overline{x^2} + 2bn\overline{x} = 0$$

$$\frac{\partial(\mathrm{SE})}{\partial m} = -2n\,\overline{xy} + 2mn\overline{x^2} + 2bn\overline{x} = 0$$

$$= -\overline{xy} + m\overline{x^2} + b\overline{x} = 0$$

$$m\overline{x^2} + b\overline{x} = \overline{x\,y}$$

$$m\frac{\overline{x^2}}{\overline{x}} + b = \frac{\overline{x\,y}}{\overline{x}}$$

one point $\left(\frac{\overline{x^2}}{\overline{x}}, \frac{\overline{x\,y}}{\overline{x}}\right)$

$$\mathrm{SE}=n\,\overline{y^{2}}-2mn\,\overline{x\,y}-2bn\,\overline{y}+m^{2}n\,\overline{x^{2}}+2mbn\,\overline{x}+nb^{2}$$

$$\frac{\partial(\mathrm{SE})}{\partial b}=-2n\,\overline{y}+2mn\,\overline{x}+2nb=0$$

$$=-\overline{y}+m\overline{x}+b=0$$

$$\overline{y}=m\overline{x}+b$$

another point $(\overline{x}, \overline{y})$

Least Squares Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$