Multiple Linear Regression

More than one independent variables are involved to express a single dependent variable.

Attempt is to increase the accuracy of the estimates.

It is expressed as,

$$y = a + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_k x_k + \varepsilon$$

where, $x_1, x_2, \ldots\ldots\ldots, x_k$ are independent variables and

y is the dependent variable.

$b_1, b_2, \ldots\ldots\ldots, b_k$ are the regression coefficients which are to be estimated.

ε is the error term

Broad steps involved in developing a linear multiple regression model

1. Hypothesize the form of the model. This involves the choice of the independent variables to be included in the model.

2. Estimate the unknown parameters, a, b_1, b_2, ….b_k.

3. Make inferences on the estimates.

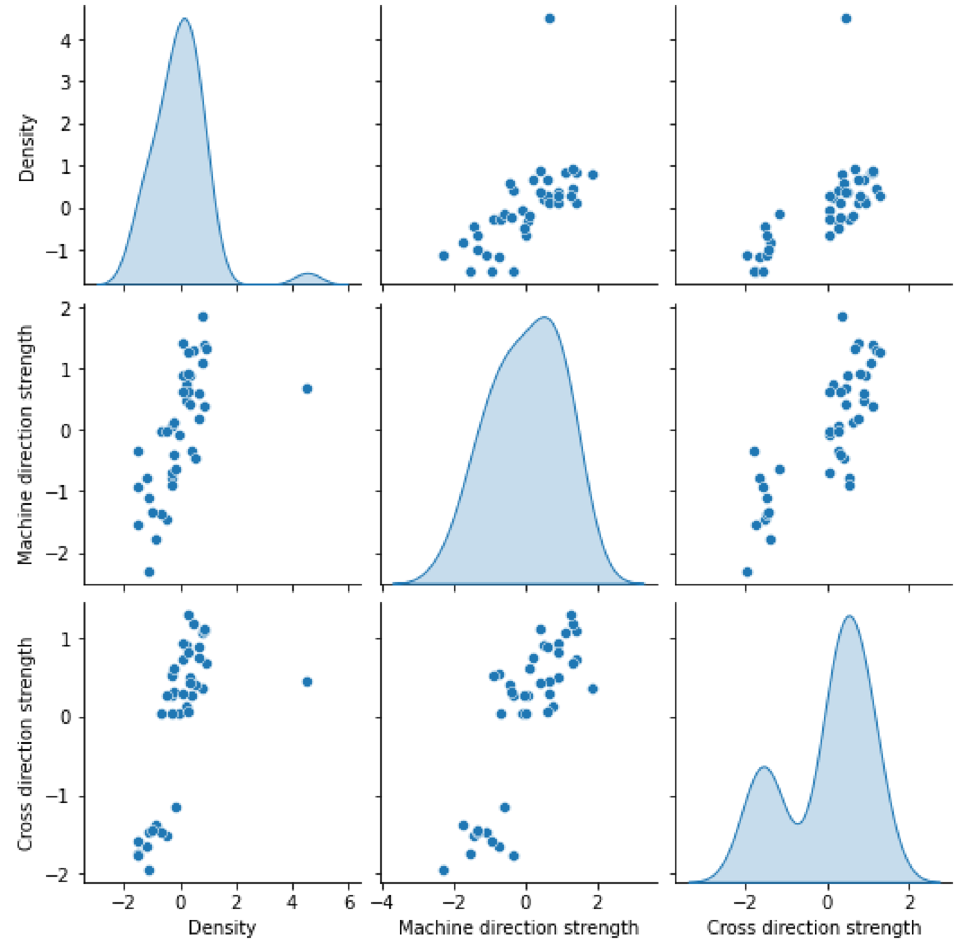
Scatterplot Matrix (SPLOM)

- Scatterplot Matrix plots all possible combinations of two or more numeric variables against one another.
- The plots are arranged in rows and columns, with the same number of rows and columns as there are variables. The point of the plot is simple. When you have many variables to plot against each other in scatterplots, it is logical to arrange the plots in rows and columns using a common vertical scale for all plots within a row (and a common horizontal scale within columns). All complete x-y pairs within each plot are used; that is pairwise deletion is used for missing data.

	Density	Machine direction strength	Cross direction strength
0	0.801	121.41	70.42
1	0.824	127.70	72.47
2	0.841	129.20	78.20
3	0.816	131.80	74.89
4	0.840	135.10	71.21

```
from scipy import stats
dfz = stats.zscore(df)
```

```
import seaborn
seaborn.pairplot(dfz,
kind='scatter',diag_kind="kde",palette="deep")
```



LEAST SQUARES ESTIMATION OF THE PARAMETERS

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

The least squares function is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

We want to minimize L with respect to $\beta_0, \beta_1, \ldots, \beta_k$. The **least squares estimates** of $\beta_0, \beta_1, \ldots, \beta_k$ must satisfy

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

Normal Equations are written as matrix form,

$$
\begin{bmatrix}
n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\
\sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}\boldsymbol{x}_{ik} \\
\vdots & \vdots & \vdots & & \vdots \\
\sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2
\end{bmatrix}
\begin{bmatrix}
\hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k
\end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i
\end{bmatrix}
$$

Coefficient of Determination

- R^2 : Proportion of variation of values of y explained by the regression model.
- $0 \leq R^2 \leq 1$
- $R^2 = 1$, indicates the regression line is a perfect estimation of linear relationship between x & y.
- $R^2 = 0$, indicates no relationship

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{SS_E / (n-p)}{SS_T / (n-1)}$$

Hypothesis testing in Multiple Linear Regression

I. Test for significance of regression

This test for significance is to determine whether a linear relationship exists between the response variable y and a set of the regressor variables x_1, x_2, ….x_k.

The hypothesis are

Ho : $b_1 = b_2 = \ldots. = b_k = 0$

H1 : $b_j \neq 0$ for at least one j.

ANOVA for testing significance of regression

Source of Variation	Sum of Squares	df	Mean Sum of Squares	F_o	*p*-value
Regression Error	SSR SSE	k n-k-1	MSR=SSR/k MSE=SSE/(n-k-1)	MSR/MSE	
Total	TSS	n-1			

n is the number of data points in the sample

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSR = TSS - SSE$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

From table we get, $F_{\alpha, k, n-k-1} = F_{table}$

If $F_o > F_{table}$, then reject H_o

OR,

if p-value $< \alpha$, then reject H_o.

II. Tests on individual regression co-efficients.

Such tests are useful in determining the potential of each of the regressor variables in the regression model.

The model might be more effective with the inclusion of an additional variable or perhaps the deletion of one or more of the regressors present in the model.

Hypothesis:
$H_o : b_j = 0$
$H_1 : b_j \neq 0$

$$t_0 = \frac{\hat{b}_j}{se\left(\hat{b}_j\right)}$$

If $\left|t_0\right| > t_{\alpha/2,n-k-1}$, OR, if p-value < α/2 , then reject H_o.

III. Confidence Interval for dependent variable.

A 100(1-α) % CI on the dependent variable is given by

$$\hat{y} - t_{\alpha/2,n-k-1}\, se(\hat{y}) \le \hat{y} \le \hat{y} + t_{\alpha/2,n-k-1}\, se(\hat{y})$$

$$se(\hat{y}) = \sqrt{MSE} = \text{Standard error of Estimate}$$

IV. Confidence Intervals of individual regression co-efficients.

A 100(1-α) % CI on the regression co-efficient b_j is given by,

$$\hat{b}_j - t_{\alpha/2, n-k-1} se\left(\hat{b}_j\right) \leq b_j \leq \hat{b}_j + t_{\alpha/2, n-k-1} se\left(\hat{b}_j\right)$$

Standardized regression coefficient

Regression model is estimated using standardized data.

Dimensionless regression co-efficient can help to compare the relative importance of each variable.

If $\left|\hat{b}_j\right| > \left|\hat{b}_i\right|$, then we can say that regressor x_j produces a larger effect than the regressor x_i.

EX 1 : The data shown in Table represent the thrust of a jet-turbine engine (y) and six candidate regressors: $x1$ = primary speed of rotation, $x2$ = secondary speed of rotation, $x3$ = fuel low rate, $x4$ = pressure, $x5$ = exhaust temperature, and $x6$ = ambient temperature at time of test.

(a) Fit a multiple linear regression model with the above data and interpret the results.
(b) Fit a multiple linear regression model using $x3$ = fuel low rate, $x4$ = pressure, and $x5$ = exhaust temperature as the regressors, and interpret the results.
(c) Refit the model using $y* = \ln(y)$ as the response variable and $x3^* = \ln(x3)$ as the regressor (along with $x4$ and $x5$). How do you compare with the previous fitted regression model?

Obs	y	x1	x2	x3	x4	x5	x6
1	4540	2140	20640	30250	205	1732	99
2	4315	2016	20280	30010	195	1697	100
3	4095	1905	19860	29780	184	1662	97
4	3650	1675	18980	29330	164	1598	97
5	3200	1474	18100	28960	144	1541	97
6	4833	2239	20740	30083	216	1709	87
7	4617	2120	20305	29831	206	1669	87
8	4340	1990	19961	29604	196	1640	87
9	3820	1702	18916	29088	171	1572	85
10	3368	1487	18012	28675	149	1522	85
11	4445	2107	20520	30120	195	1740	101
12	4188	1973	20130	29920	190	1711	100
13	3981	1864	19780	29720	180	1682	100
14	3622	1674	19020	29370	161	1630	100
15	3125	1440	18030	28940	139	1572	101
16	4560	2165	20680	30160	208	1704	98
17	4340	2048	20340	29960	199	1679	96
18	4115	1916	19860	29710	187	1642	94
19	3630	1658	18950	29250	164	1576	94
20	3210	1489	18700	28890	145	1528	94

21	4330	2062	20500	30190	193	1748	101
22	4119	1929	20050	29960	183	1713	100
23	3891	1815	19680	29770	173	1684	100
24	3467	1595	18890	29360	153	1624	99
25	3045	1400	17870	28960	134	1569	100
26	4411	2047	20540	30160	193	1746	99
27	4203	1935	20160	29940	184	1714	99
28	3968	1807	19750	29760	173	1679	99
29	3531	1591	18890	29350	153	1621	99
30	3074	1388	17870	28910	133	1561	99
31	4350	2071	20460	30180	198	1729	102
32	4128	1944	20010	29940	186	1692	101
33	3940	1831	19640	29750	178	1667	101
34	3480	1612	18710	29360	156	1609	101
35	3064	1410	17780	28900	136	1552	101
36	4402	2066	20520	30170	197	1758	100
37	4180	1954	20150	29950	188	1729	99
38	3973	1835	19750	29740	178	1690	99
39	3530	1616	18850	29320	156	1616	99
40	3080	1407	17910	28910	137	1569	100

EX 2 Patient Satisfaction Data

The regressor variables are the patient's age, an illness severity index (higher values indicate greater severity), an indicator variable denoting whether the patient is a medical patient (0) or a surgical patient (1), and an anxiety index (higher values indicate greater anxiety).

Fit a multiple linear regression model to the satisfaction response using age, illness severity, and the anxiety index as the regressors.

Observation	Age	Severity	Surg-Med	Anxiety	Satisfaction
1	55	50	0	2.1	68
2	46	24	1	2.8	77
3	30	46	1	3.3	96
4	35	48	1	4.5	80
5	59	58	0	2.0	43
6	61	60	0	5.1	44
7	74	65	1	5.5	26
8	38	42	1	3.2	88
9	27	42	0	3.1	75
10	51	50	1	2.4	57
11	53	38	1	2.2	56
12	41	30	0	2.1	88
13	37	31	0	1.9	88
14	24	34	0	3.1	102
15	42	30	0	3.0	88
16	50	48	1	4.2	70
17	58	61	1	4.6	52
18	60	71	1	5.3	43
19	62	62	0	7.2	46
20	68	38	0	7.8	56
21	70	41	1	7.0	59
22	79	66	1	6.2	26
23	63	31	1	4.1	52
24	39	42	0	3.5	83
25	49	40	1	2.1	75

```
import pandas as pd
import statsmodels.api as sm
import numpy as np

df = pd.read_csv("C:/Users/ … /7Engine.csv")

# PROVIDING DATA

X = df.iloc[:,2:].copy()
y = df['y'].copy()

X, y = np.array(X), np.array(y)
X = sm.add_constant(X)

model1 = sm.OLS(y, X)
results1 = model1.fit()
print(results1.summary())
print("\n Fitted Values:\n")
y_pred = results1.fittedvalues.round(2)
print(y_pred)
```