STATISTICS IS THE GRAMMAR OF SCIENCE

PROBABILITY AND STATISTICS

LECTURE – 23

REGRESSION ANALYSIS

MULTIPLE LINEAR REGRESSION MODEL

PREPARED BY
HAZBER SAMSON
FAST NUCES ISLAMABAD

REGRESSION ANALYSIS

INTRODUCION: At the beginning of the nineteenth century, **Legendre and Gauss** published papers on the *method of least squares*, which implemented the earliest form of what is now known as *linear regression*. The term *regression* was first introduced by **Francis Galton**. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or “regress” toward the average height in the population as a whole.1 In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton’s *law of universal regression* was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups.2 He found that the average height of sons of a group of tall fathers was less than their fathers’ height and the average height of sons of a group of short fathers was greater than their fathers’ height, thus “regressing” tall and short sons alike toward the average height of all men. In the words of Galton, this was “regression to mediocrity. The modern interpretations of regression is, as follows however,

DEFINITION: Regression analysis is a statistical technique for investigating and modeling the relationship between variables.

Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

Usually, Researchers want more than just finding the best linear approximation of one variable given a set of others. They want valid statistical relationships. They want to draw conclusions about what happens if one of the variables actually changes. That is: they want to say something about things that are not observed (yet). In this case, they want to find a fundamental relationship. To do this it is assumed that there is a general relationship that is valid for all possible observations from a well-defined population. Restricting attention to linear relationships, we specify a statistical model as linear regression model. Some examples of analyses using regression models include the following:

- Estimating weight gain by the addition to children’s diet of different amounts of various dietary supplements
- Predicting scholastic success (grade point ratio) based on students’ scores on an aptitude or entrance test
- Estimating amounts of sales associated with levels of expenditures for various types of advertising
- Predicting fuel consumption for home heating based on daily temperatures and other weather factors
- Estimating changes in interest rates associated with the amount of deficit spending

Regression analysis is used to predict the value of one variable on the basis of the other variables. Regression analysis is generally classified into two kinds, **simple** and **multiple**. The regression will be **simple** if it involves only one independent variable and it will be **multiple** if it involves more than one independent variables.

MULTIPLE LINEAR REGRESSION MODEL (MLRM)

The dependence of one variable upon more than one independent variables is known as Multiple Regression. Generally in real life scenarios we consider more than one explanatory variables for example

- Predicting GPA of students based on aptitude test, high school grades and IQ level.
- Estimating changes in sales associated with advertisement, Number of employees Working Hours and Salaries of Employees.

MULTIPLE LINEAR REGRESSION MODEL

A mathematical expression that shows the dependence of one variable upon more than one independent variables is known as Multiple Linear Regression Model.

The standard form of Multiple linear regression Model can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

Where

- $Y = Dependent\ Variable$
- $\beta_0 = Intercept$
- $\beta_1, \beta_2, ..., \beta_k = Partial\ \mathrm{Re}\, gression\ Coefficients$
- $X_1, X_2, ..., X_k = Independent\ Variables$
- $\beta_0, \beta_1, \beta_2, ..., \beta_k = Parameters\, of\ the\, Model$
- $\varepsilon = Error Term$

In general if we have n observations for the above MLRM then it can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_j X_{ij} ... + \beta_k X_{ik} + \varepsilon_i$$

Here X_{ij} represents i^{th} observation of the j^{th} independent variable. Where
$i = 1, 2, ..., n \quad and \quad j = 1, 2, ..., k$

If we expand the above Model it becomes

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + ... + \beta_k X_{1k} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + ... + \beta_k X_{2k} + \varepsilon_2$$

. . . . . .
. . . . . .
. . . . . .

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + ... + \beta_k X_{nk} + \varepsilon_n$$

ASSUMPTIONS OF CLASSICAL LINEAR REGRESSION MODEL

Consider the classical linear regression model (CLRM)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

ASSUMPTION-1 ***LINEARITY*** *The regression model is linear in the parameters,* though it may or may not be linear in the variables. It means that the dependent variable can be calculated as a linear function of a specific set of independent variables, plus a disturbance term.

ASSUMPTION-2 ***NUMBER OF OBSERVATIONS*** The number of observations *n* must be greater than the number of parameters to be estimated: Alternatively, the number of observations must be greater than the number of explanatory variables.

ASSUMPTION-3 INDEPENDENT VARIABLES HAVE SOME VARIATION
By this assumption we mean that not all observations of any independent variable X are the same; at least one has to be different so that the sample Var(X) is not 0. Furthermore, there can be no outliers in the values of the *X* variable.

ASSUMPTION-4 ***EXPECTED VALUE OF THE ERROR TERM IS ZERO*** $E(\varepsilon_i) = 0 \ \ for\, all\, i.$

Mean value of error terms is zero. ie This means that the disturbance is a genuine disturbance, so if we took a large number of samples the mean disturbance will be zero.

ASSUMPTION-5 ***HOMOSCEDASTICITY OF ERROR TERMS*** $Var(\varepsilon_i) = \sigma^2 \ \ for\, all\, i.$

Variance of error terms is same for all observations. This means that all error terms have the same variance regardless of the value of X.

ASSUMPTION-6 ***INDEPENDENCE OF ERROR TERMS*** $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \ for\, all\, i \neq j$

Error terms are independent of each other. So there is no correlation between two error terms. In short, the observations are sampled independently.

ASSUMPTION-7 FIXED VALUES OF INDEPENDENT VARIABLES $Cov\,(X_i, \varepsilon_i) = 0 \ \ for\, all\, i\,.$

Values taken by any regressor *X* may be considered fixed in repeated samples also X variable(s) and the error terms are independent of each other.

ASSUMPTION-8 NO EXACT LINEAR RELATIONSHIP BETWEEN REGRESSORS
There are no exact linear relationships among the sample values of any two or more of the explanatory variables.

ASSUMPTION-9 ***NORMALITY OF ERROR TERMS*** $\varepsilon_i \sim N(0, \sigma^2) \ \forall \ i$

Error terms are normally distributed with mean zero and variance σ^2

ASSUMPTION-10 ***NO SPECIFICATION BIAS***
The model is correctly specified.

MULTIPLE LINEAR REGRESSION MODEL WITH TWO REGRESSORS

The standard form of MLRM with two regressors can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

For estimation consider the model $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

NORMAL EQUATIONS

- $\sum Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_1 + \hat{\beta}_2 \sum X_2$
- $\sum X_1 Y = \hat{\beta}_0 \sum X_1 + \hat{\beta}_1 \sum X_1{}^2 + \hat{\beta}_2 \sum X_1 X_2$
- $\sum X_2 Y = \hat{\beta}_0 \sum X_2 + \hat{\beta}_1 \sum X_1 X_2 + \hat{\beta}_2 \sum X_2{}^2$

METHOD-1 MATRICES APPROACH

Consider the system of Normal Equations

$$\sum Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_1 + \hat{\beta}_2 \sum X_2 \qquad (1)$$

$$\sum X_1 Y = \hat{\beta}_0 \sum X_1 + \hat{\beta}_1 \sum X_1{}^2 + \hat{\beta}_2 \sum X_1 X_2 \qquad (2)$$

$$\sum X_2 Y = \hat{\beta}_0 \sum X_2 + \hat{\beta}_1 \sum X_1 X_2 + \hat{\beta}_2 \sum X_2{}^2 \qquad (3)$$

In Matrix Form we have

$$\begin{pmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X & \sum X_2^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X & \sum X_2^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \end{pmatrix}$$

We shall find the inverse of the Matrix of Coefficients and then multiply it by the column matrix and get the values of estimators of the multiple linear regression model.

METHOD-2 DEVIATION FORM For estimation consider $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \beta_2 X_2$

In deviation form $Y = \beta_1 x_1 + \beta_2 x_2 \;\; where \; x_1 = X_1 - \overline{X}_1 \;\; and \;\; x_2 = X_2 - \overline{X}_2$

$$\hat{\beta}_1 = \frac{\left(\sum x_1 y\right)\left(\sum x_2{}^2\right) - \left(\sum x_2 y\right)\left(\sum x_1 x_2\right)}{\left(\sum x_1{}^2\right)\left(\sum x_2{}^2\right) - \left(\sum x_1 x_2\right)^2}$$

$$\hat{\beta}_2 = \frac{\left(\sum x_2 y\right)\left(\sum x_1{}^2\right) - \left(\sum x_1 y\right)\left(\sum x_1 x_2\right)}{\left(\sum x_1{}^2\right)\left(\sum x_2{}^2\right) - \left(\sum x_1 x_2\right)^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X_1} - \hat{\beta}_2 \overline{X_2}$$

Where

- $\sum x_1 x_2 = \sum X_1 X_2 - \frac{\left(\sum X_1\right)\left(\sum X_2\right)}{n}$
- $\sum x_1 y = \sum X_1 Y - \frac{\left(\sum X_1\right)\left(\sum Y\right)}{n}$
- $\sum x_2 y = \sum X_2 Y - \frac{\left(\sum X_2\right)\left(\sum Y\right)}{n}$
- $\sum x_1{}^2 = \sum X_1{}^2 - \frac{\left(\sum X_1\right)^2}{n}$
- $\sum x_2{}^2 = \sum X_2{}^2 - \frac{\left(\sum X_2\right)^2}{n}$

MULTIPLE STANDARD ERROR OF ESTIMATE

$$S_{Y.23} = \sqrt{\frac{\sum\left(Y - \widehat{Y}\right)^2}{n-3}}$$

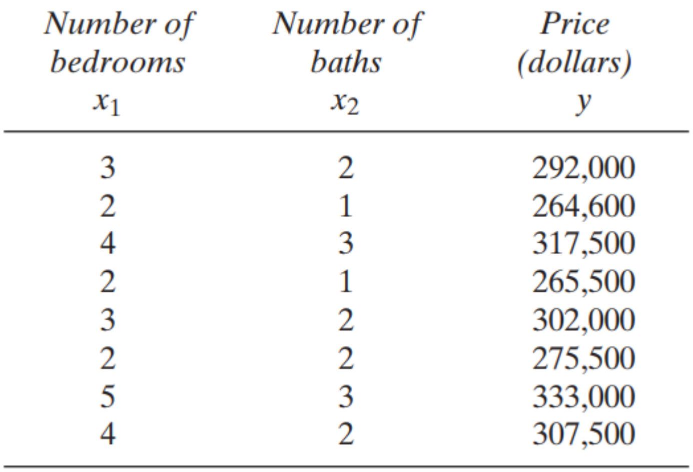
$$S_{Y.23} = \sqrt{\frac{\sum Y^2 - \beta_0 \sum Y - \beta_1 \sum X_1 Y - \beta_2 \sum X_2 Y}{n-3}}$$

CO-EFFICIENT OF MULTIPLE DETERMINATION

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST} = \frac{\sum\left(\widehat{Y} - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

$$R^2 = 1 - \frac{Un\exp lained\,Variation}{Total\,Variation} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum\left(Y - \widehat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

EXAMPLE-1 The following data show the number of bedrooms, the number of baths, and the prices at which a random sample of eight one-family houses sold in a certain large housing development:



Number of bedrooms x_1	*Number of baths* x_2	*Price (dollars)* y
3	2	292,000
2	1	264,600
4	3	317,500
2	1	265,500
3	2	302,000
2	2	275,500
5	3	333,000
4	2	307,500

(a) Use the method of least squares to fit a linear equation that will enable us to predict the average sales price of a one-family house in the given housing development in terms of the number of bedrooms and the number of baths. Also Interpret the Results.
(b) Predict the sales price of a three-bedroom house with two baths in the subject housing development.
(c) Find Coefficient of Determination and Interpret it.

SOLUTION (a) We have multiple linear regression model given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

For estimation consider the model

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Normal Equations are given by

$$\sum Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_1 + \hat{\beta}_2 \sum X_2 \qquad (1)$$
$$\sum X_1 Y = \hat{\beta}_0 \sum X_1 + \hat{\beta}_1 \sum X_1{}^2 + \hat{\beta}_2 \sum X_1 X_2 \qquad (2)$$
$$\sum X_2 Y = \hat{\beta}_0 \sum X_2 + \hat{\beta}_1 \sum X_1 X_2 + \hat{\beta}_2 \sum X_2{}^2 \qquad (3)$$

In Matrix Form we have

$$\begin{pmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X & \sum X_2^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \end{pmatrix}$$

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix} n & \sum X_1 & \sum X_2\\ \sum X_1 & \sum X_1^2 & \sum X_1X_2\\ \sum X_2 & \sum X_1X & \sum X_2^2\end{pmatrix}^{-1}\begin{pmatrix}\sum Y\\ \sum X_1Y\\ \sum X_2Y\end{pmatrix} \qquad (A)$$

Let’s find the values of the entries of the matrices

Y	X_1	X_2	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
292000	3	2	9	4	6	876000	584000
264600	2	1	4	1	2	529200	264600
317500	4	3	16	9	12	1270000	952500
265500	2	1	4	1	2	531000	265500
302000	3	2	9	4	6	906000	604000
275500	2	2	4	4	4	551000	551000
333000	5	3	25	9	15	1665000	999000
307500	4	2	16	4	8	1230000	615000
2357600	**25**	**16**	**87**	**36**	**55**	**7558200**	**4835600**

Substituting these values in the above normal equations we get

$$2{,}357{,}600 = 8\hat{\beta}_0 + 25\hat{\beta}_1 + 16\hat{\beta}_2$$

$$7{,}558{,}200 = 25\hat{\beta}_0 + 87\hat{\beta}_1 + 55\hat{\beta}_2$$

$$4{,}835{,}600 = 16\hat{\beta}_0 + 55\hat{\beta}_1 + 36\hat{\beta}_2$$

Substituting these values in equation (A) we get

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix}8 & 25 & 16\\ 25 & 87 & 55\\ 16 & 55 & 36\end{pmatrix}^{-1}\begin{pmatrix}2357600\\ 7558200\\ 4835600\end{pmatrix}=\begin{pmatrix}1.2738 & -0.238 & -0.202\\ -0.238 & 0.3809 & -0.476\\ -0.202 & -0.476 & 0.8452\end{pmatrix}^{-1}\begin{pmatrix}2357600\\ 7558200\\ 4835600\end{pmatrix}=\begin{pmatrix}224928.57\\ 15314.28\\ 10957.14\end{pmatrix}$$

So the fitted regression model is given by $\hat{Y} = 224929 + 15314X_1 + 10957X_2 \quad (B)$

<u>INTERPRETATION OF RESULTS OF THE MODEL</u>

Interpretation of $\hat{\beta}_0 = 224929$ **:** If a house has no bedrooms and no baths, even then its price will be $\$224929$.

Interpretation of $\hat{\beta}_1 = 15314$: If number of bedrooms in a house is increased by one then its price will be increased by $\$15314$ on average in the long run, keeping other factors as constant.

Interpretation of $\hat{\beta}_2 = 10957$: If number of baths in a house is increased by one then its price will be increased by $\$10957$ on average in the long run, keeping other factors as constant.

(b) Prediction

We have to predict the sales price of a three-bedroom house with two baths in the subject housing development. i.e. $\hat{Y}=?,\ X_1=3\ \ and\ \ X_2=2$

Put the values in equation (B) we get

$$\hat{Y}=224929+15314(3)+10957(2)=\$292785$$

(b) Coefficient of Determination

Coefficient of Determination is given by

$$R^2=\frac{Explained\ Variation}{Total\ Variation}=\frac{SSR}{SST}=\frac{\sum\left(\hat{Y}-\overline{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2}$$

here $\overline{Y}=\frac{\sum Y}{n}=\frac{2357600}{8}=294700$

Y	$\hat{Y}$	$\hat{Y}-\overline{Y}$	$(\hat{Y}-\overline{Y})^2$	$Y-\overline{Y}$	$(Y-\overline{Y})^2$
292000	292785	-1915	3667225	-2700	7290000
264600	266514	-28186	794450596	-30100	906010000
317500	319056	24356	593214736	22800	519840000
265500	266514	-28186	794450596	-29200	852640000
302000	292785	-1915	3667225	7300	53290000
275500	277471	-17229	296838441	-19200	368640000
333000	334370	39670	1573708900	38300	1466890000
307500	308099	13399	179533201	12800	163840000
2357600	**2357594**	**-6**	**4239530920**	**0**	**4338440000**

$$R^2=\frac{\sum\left(\hat{Y}-\overline{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2}=\frac{4239530920}{4338440000}=0.98$$

Interpretation: $R^2=0.98$ shows that 98% variation in Price of a house is explained due to number of bedrooms and number of baths while remaining 2% is due to other factors.

EXERCISE – 5.3

MULTIPLE LINEAR REGRESSION MODEL

1. The following are sample data provided by a moving company on the weights of six shipments, the distances they were moved, and the damage that was incurred:

Weight (1,000 *pounds*) x_1	*Distance* (1,000 *miles*) x_2	*Damage* (*dollars*) y
4.0	1.5	160
3.0	2.2	112
1.6	1.0	69
1.2	2.0	90
3.4	0.8	123
4.8	1.6	186

(a) Fit a multiple linear regression model using the above data set also interpret the results.
(b) Estimate the damage when a shipment weighing 2,400 pounds is moved 1,200 miles.
(c) Find Coefficient of Determination and interpret it.

$(Ans:(a)\hat{y}=14.56+30.11x_1+12.16x_2\ \ (b)\,\hat{y}=\$101.41\ \ (c)\,R^2=0.9197)$

2. The following are data on the average weekly profits (in $1,000) of five restaurants, their seating capacities, and the average daily traffic (in thousands of cars) that passes their locations:

Seating capacity x_1	*Traffic count* x_2	*Weekly net profit* y
120	19	23.8
200	8	24.2
150	12	22.0
180	15	26.2
240	16	33.5

(a) Fit a multiple linear regression model using the above data set also interpret the results.
(b) Predict the average weekly net profit of a restaurant with a seating capacity of 210 at a location where the daily traffic count averages 14,000 cars.
(c) Find Coefficient of Determination and interpret it.

$(Ans:(a)\hat{y}=-0.627+0.097x_1+0.662x_2\ \ (b)\,\hat{y}=84\ \ (c)\,R^2=0.9943)$

3. The following data represent the chemistry grades for a random sample of 12 freshmen at a certain college along with their scores on an intelligence test administered while they were still seniors in high school. we were also given the number of class periods missed by the 12 students taking the chemistry course. The complete data are shown.

Student	**Chemistry Grade, y**	**Test Score, x_1**	**Classes Missed, x_2**
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	76	50	1
12	74	55	4

(a) Fit a multiple linear regression model using the above data set also interpret the results.
(b) Estimate the chemistry grade for a student who has an intelligence test score of 60 and missed 4 classes.
(c) Find Coefficient of Determination and interpret it.

$(Ans:(a)\hat{y}=27.5467+0.9217x_1+0.2842x_2\ \ (b)\,\hat{y}=84\ \ (c)\,R^2=\quad)$

4. The following are data on the percent effectiveness of a pain reliever and the amounts of three different medications (in milligrams) present in each capsule:

Medication A	*Medication B*	*Medication C*	*Percent effective*
x_1	x_2	x_3	y
15	20	10	47
15	20	20	54
15	30	10	58
15	30	20	66
30	20	10	59
30	20	20	67
30	30	10	71
30	30	20	83
45	20	10	72
45	20	20	82
45	30	10	85
45	30	20	94

(a) Fit a multiple linear regression model using the above data set also interpret the results.
(b) Predict percent effectiveness of a pain reliever if a capsule contains Medication A 12 mg, Medication B 25mg and medication C 8mg.
(c) Find Coefficient of Determination and interpret it.

$$(Ans:(a)\hat{y}=-102.71+0.61x_1+8.92x_2+1.44x_3+0.01x_4\ \ (b)\,\hat{y}=287.56\ \ (c)\,R^2=\qquad)$$

5. The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature $x1$, the number of days in the month $x2$, the average product purity $x3$, and the tons of product produced $x4$. The past year's historical data are available and are presented in the following table.

$\boldsymbol{y}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

(a) Fit a multiple linear regression model using the above data set also interpret the results.
(b) Predict power consumption for a month in which$x1$ = 75◦F, $x2$ = 24 days, $x3$ = 90%, and $x4$ = 98 tons.
(c) Find Coefficient of Determination and interpret it.

$$(Ans:(a)\hat{y}=-102.71+0.61x_1+8.92x_2+1.44x_3+0.01x_4\ \ (b)\,\hat{y}=287.56\ \ (c)\,R^2=\qquad)$$